TheMathCompany

CUSTOMER SEGMENTATION

MAJID AL FUTTAIM – NAJM & VOX

Capstone Review Phase 2

AGENDA

- Review 1 - Quick Recap
- Our Process
- Dataset Summary
- Factor Mapping and Hypothesis
- Exploratory Data Analysis
  - Data Cleaning & Imputation
  - Univariate & Bivariate Analysis
  - Correlation Matrix & VIF
- Next Steps

WHERE WE LEFT OFF

The credit card business of the company (NAJM) is interested in capitalizing untapped acquisition potential within its movie customer base (VOX)

Problem at Hand:
How to identify and acquire profitable customers for NAJM from VOX ?

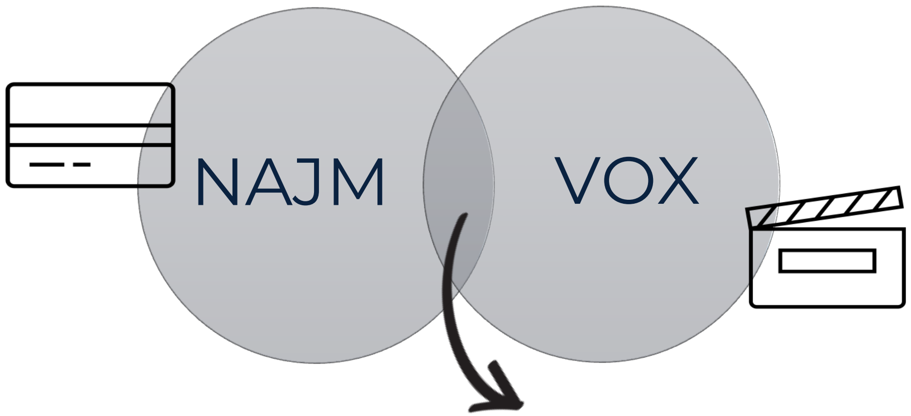
Analytics Problem

- To **understand** the **behaviour of customers** who use NAJM credit cards for payments at VOX cinemas
- To **identify profitable customers** who will purchase NAJM credit cards

Analytics Outcome

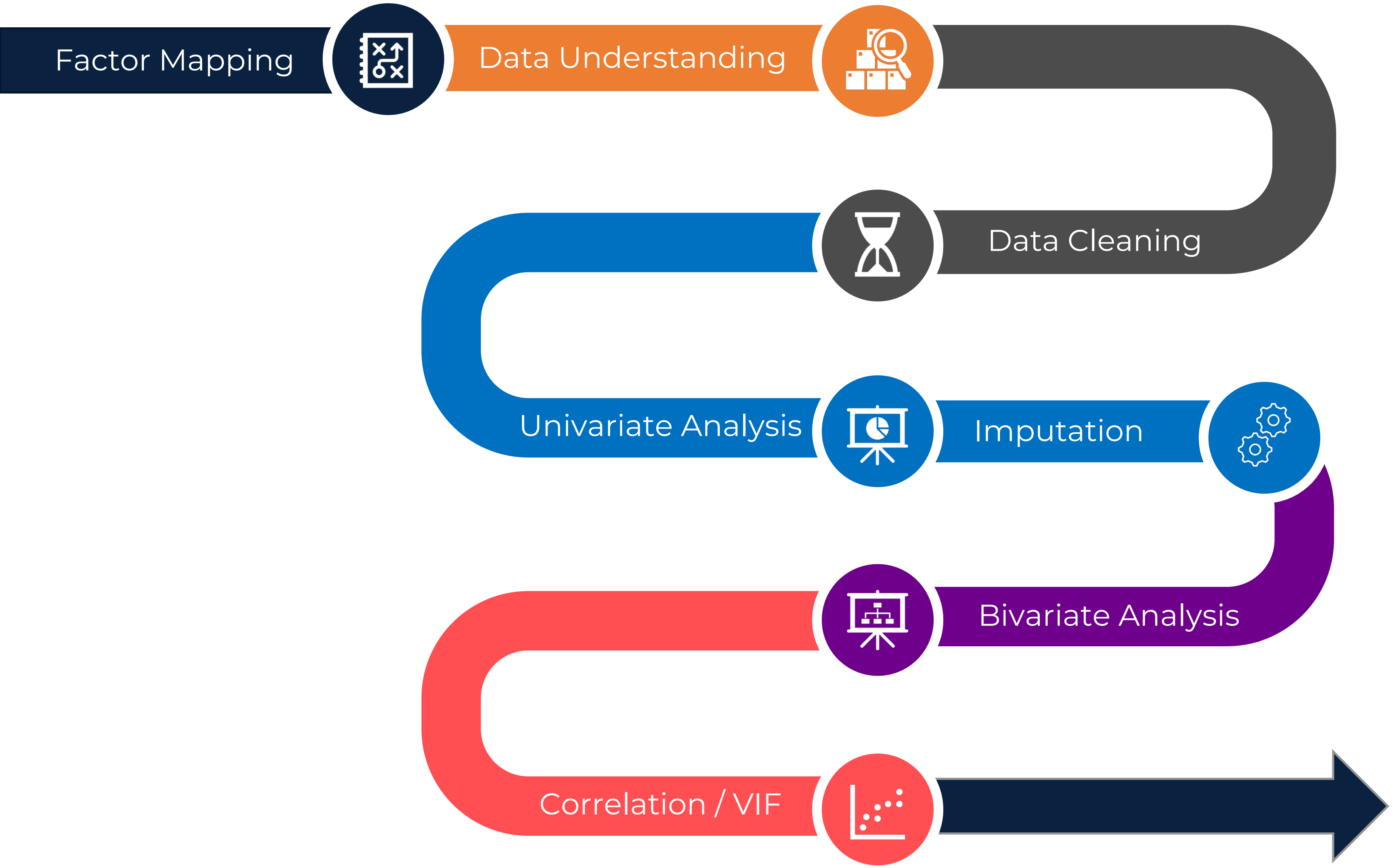
- **Characteristics** or factors with which a customer can be deemed profitable
- **Framework** to identify profitable customers to target for NAJM credit cards

Majid Al Futtaim



Customers of Interest
Customers who use NAJM credit cards for payments at VOX Cinemas

OUR PROCESS



Factor Mapping

- Brainstormed possible factors
- Framed hypothesis

Data Understanding

- Created data dictionary
- Summarized dataset

Data Cleaning

- Preliminary preprocessing

Univariate Analysis

- Distribution of data variables
- Outlier identification
- Imputed missing values

Bivariate Analysis

- Relationship b/w. data variables
- Testing hypothesis

Correlation/ VIF

- Generated correlation matrix
- VIF iterations

Final Features*

*Contingent to changes

Proprietary and Confidential

DATASET - SUMMARY

VOX customer level data

1 year

Aggregated data from 2018-2019

26

VOX cinema locations

Data

65,970 Records → **43,482** Duplicates → **22,488** Customers

Data Dictionary

53 Data Variables →

- **39** Numerical
- **12** Categorical
- **2** Time Series

→

- **10** >= 90% missing values
- **17** 1-90% missing values
- **26** < = 1% missing values

Click here for a detailed view

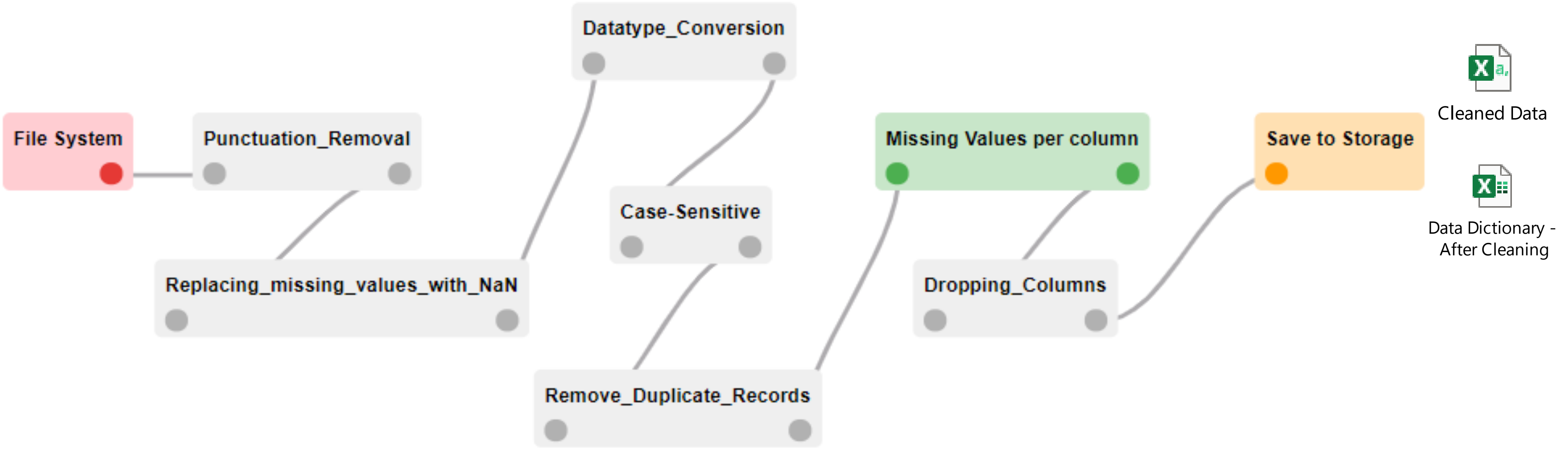
FACTOR MAPPING

Factor Mapping

Factor	Hypothesis	Conclusion
Snacks And Refreshments	People purchasing snacks and refreshments on their visits to VOX are profitable	TRUE
Spending Capacity	Customers with high spending capacity can be profitable	TRUE
Location of VOX Cinema	Location of the VOX cinema theatre affect profitability	TRUE
Offers Availed	Customers who avail offers are NOT profitable	TRUE
New VOX Users	Customers who are new to VOX can be profitable	FALSE
Quality of Screens	Customers visiting premier screens are more profitable	TRUE
Day of Visit	Customers visiting on weekends are profitable	FALSE
Total Transaction Amount	Customers with Higher overall ticket amount are profitable	FALSE
Frequency of Visits	Customers who visit VOX cinemas frequently are profitable	TRUE
# of Transactions	Customer making more transactions are profitable	NOT ENOUGH DATA
Seasonality	Customers visiting during holidays are profitable	NOT ENOUGH DATA
Cancellation	Customers who don’t apply for cancellation are more profitable	NOT ENOUGH DATA

DATA CLEANING - STEPS

CO.DX BLUEPRINT

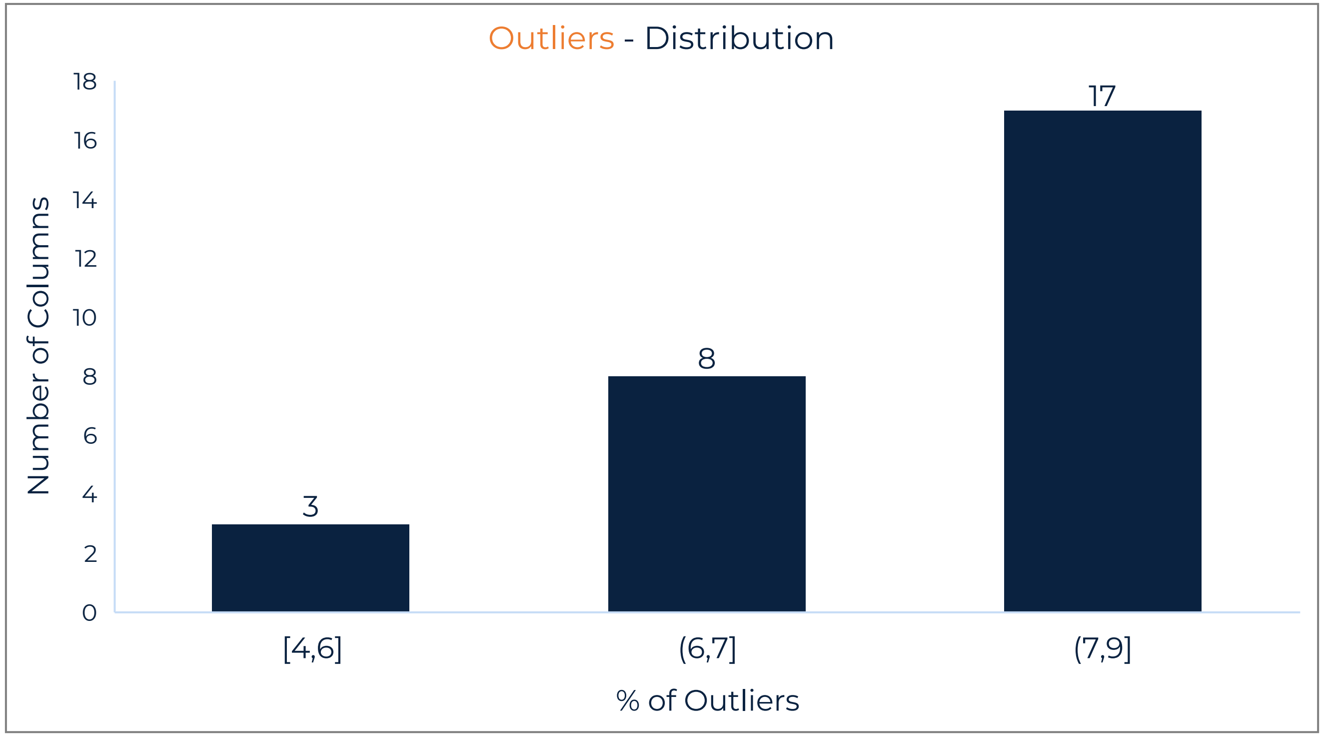


Initial Steps in Data Cleaning

- Punctuation Removal - Removal of garbage values and punctuations Ex: '60).' , '/.;2' , '?PG13/)'
- Null Value Formatting - #VALUE!, SPACE to NA
- Data type Conversion - String to Date time Format, String to Float Ex: "4362.456" to 4362.456
- Case Sensitivity - All String values changed to standard case Ex: *mall of emirates new* to *Mall of Emirates New*
- Duplicates Removal - Dropping duplicate records
- Dropping columns - Columns with >90% missing values dropped

UNIVARIATE ANALYSIS

OUTLIER IDENTIFICATION



Observations:

- Minimum number of data points so we just identified outliers rather than treating them
- Outlier treatment would manipulate the pre-existing data which will affect our model performance

DATA IMPUTATION - PROCESS

Numerical
< =1% missing – 22 / 39 columns

Imputed
with Median

Categorical
< =1% missing - 6 / 12 columns

Imputed
with Mode

Numerical
1-90% missing – 8 / 39 columns

Imputed
with KNN

Categorical
50 - 90% missing
Flags - 4 / 12 columns

Imputed
with other columns

Numerical - 9 / 39 columns
> 90% missing & can't be imputed with other columns

Dropped Columns

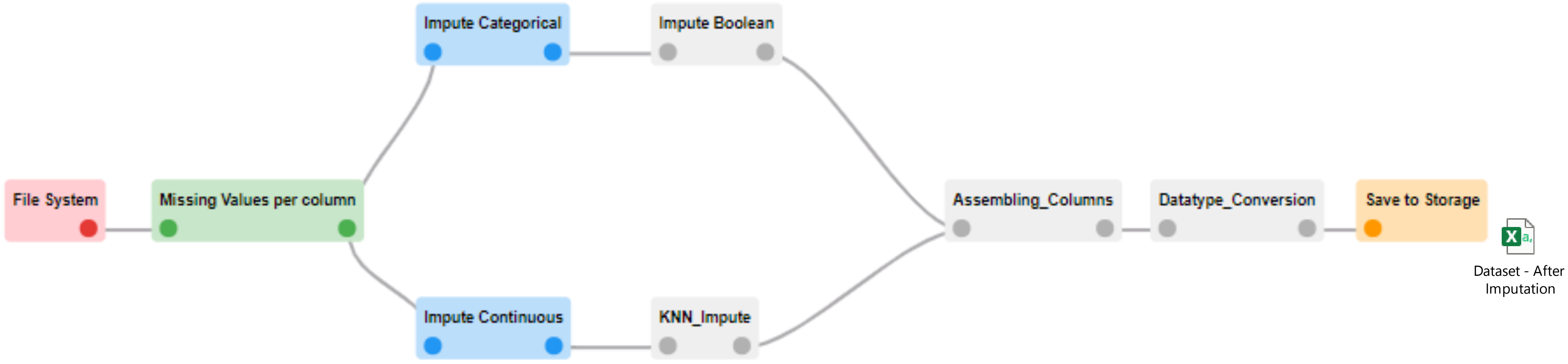
Categorical
> 90% missing – 1 / 12 columns
(New Customer* column)

Imputed
with First Transaction Date

*New customer : First transaction between Jan. 2018 – Dec. 2019

DATA IMPUTATION - STEPS

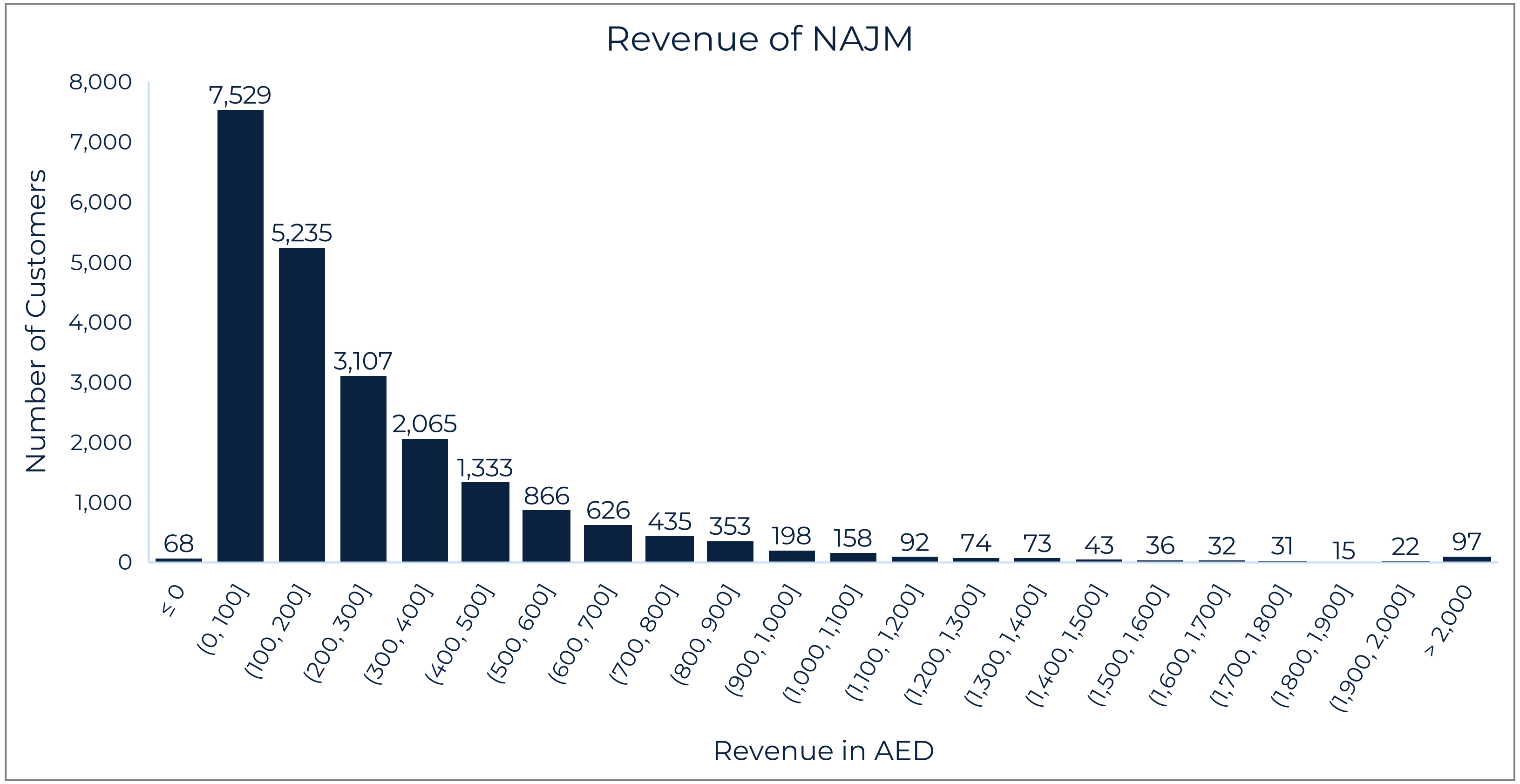
CO.DX BLUEPRINT



Steps in Data Imputation

- Categorical Values with null values <1% are imputed with their Mode
- Numeric variables with null values <1% are imputed with their median value
- Numeric variables with null values >90% are dropped definitely
- 1-90% missing Numeric variables are imputed with KNN imputation

REVENUE OF NAJM AND PROFITABILITY



Univariate Analysis

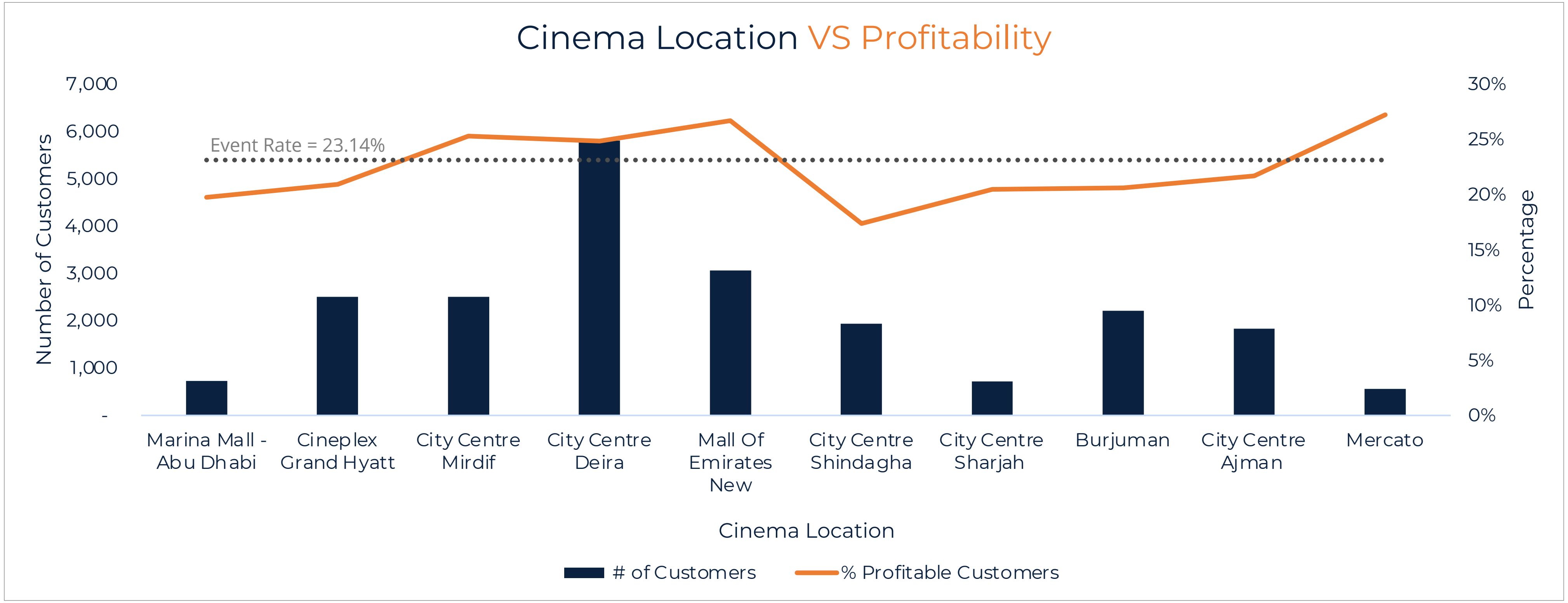
Observations:

- 68 customers generate negative revenue
- ~57% of the VOX customers generate a revenue less than 200 AED
- The threshold above which a customer is deemed profitable is > = AED 350
  - 23.14% of the customers are profitable

BIVARIATE ANALYSIS

LOCATION OF VOX CINEMA AFFECTS PROFITABILITY

Bivariate Analysis



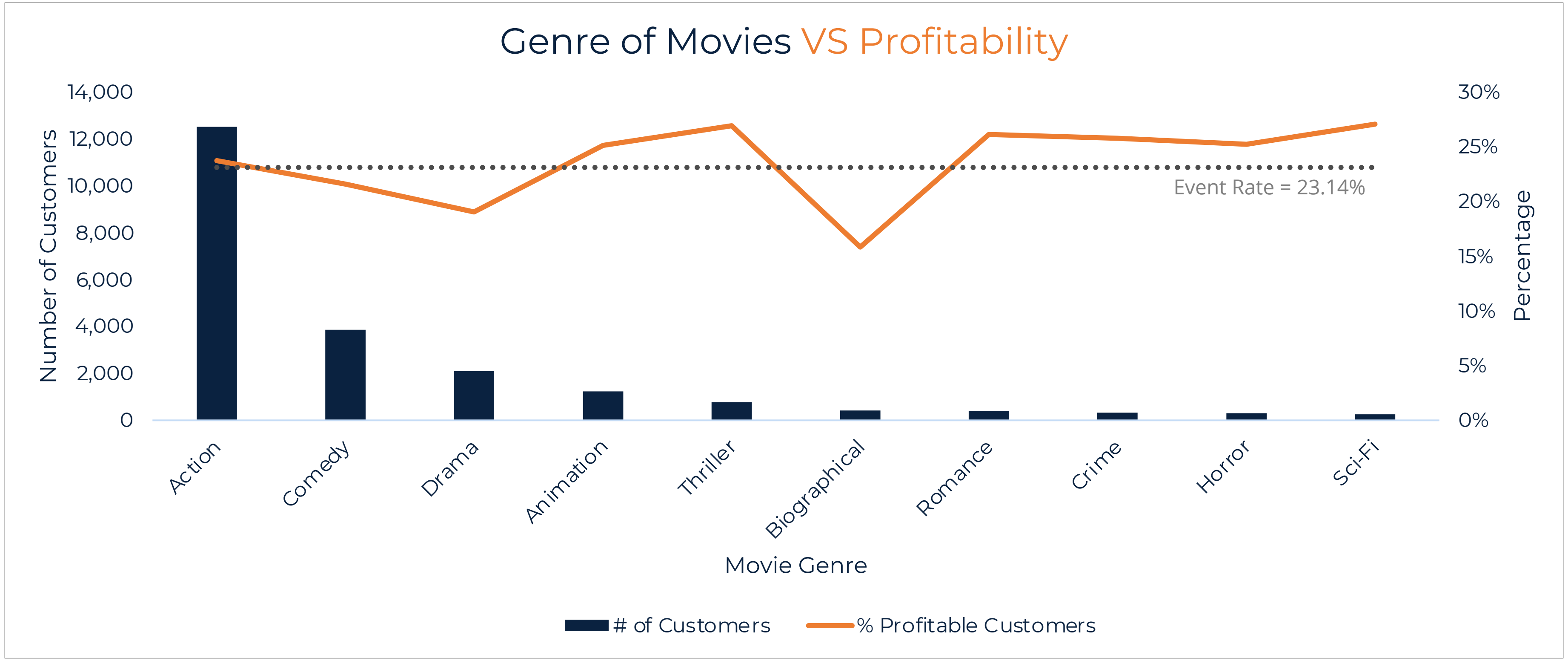
VOX Ticket Prices

Observations:

- ~26% of the customers visit City Centre Deira, the highest compared to other locations such as City Centre -Mirdif and Mall of Emirates New, yet they all attract highly profitable customers

GENRE WATCHED AFFECTS PROFITABILITY

Bivariate Analysis

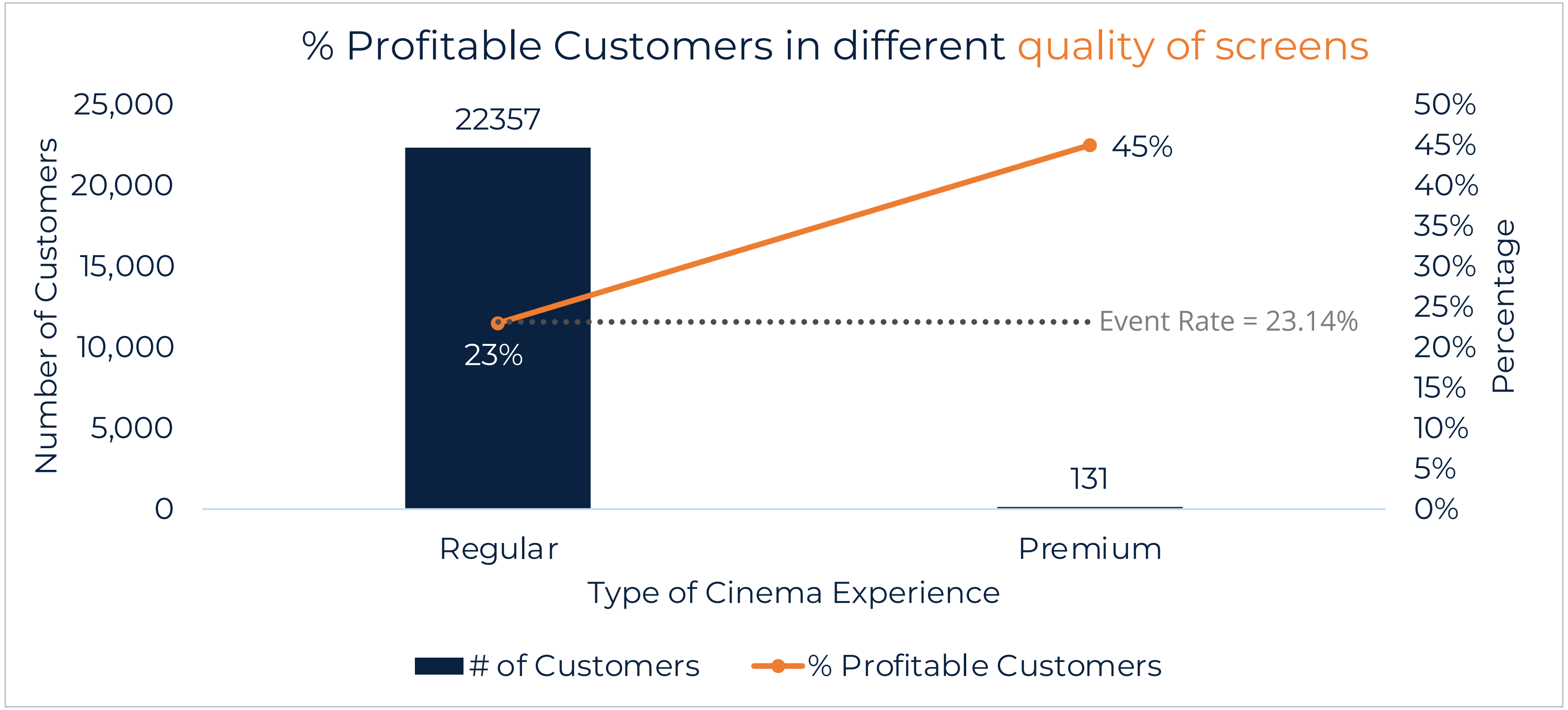


VOX Ticket Prices

Observations:

- ~55% of the customers prefer watching Action movies
- Genres like Animation, Thriller, Romance, Sci-Fi have relatively high profitability even though the # of customers visited in relatively less – WHY? - These movies are screened in premium experience where standard tickets cost around AED 80, while the regular experience tickets cost around AED 40

QUALITY OF SCREEN AFFECTS PROFITABILITY



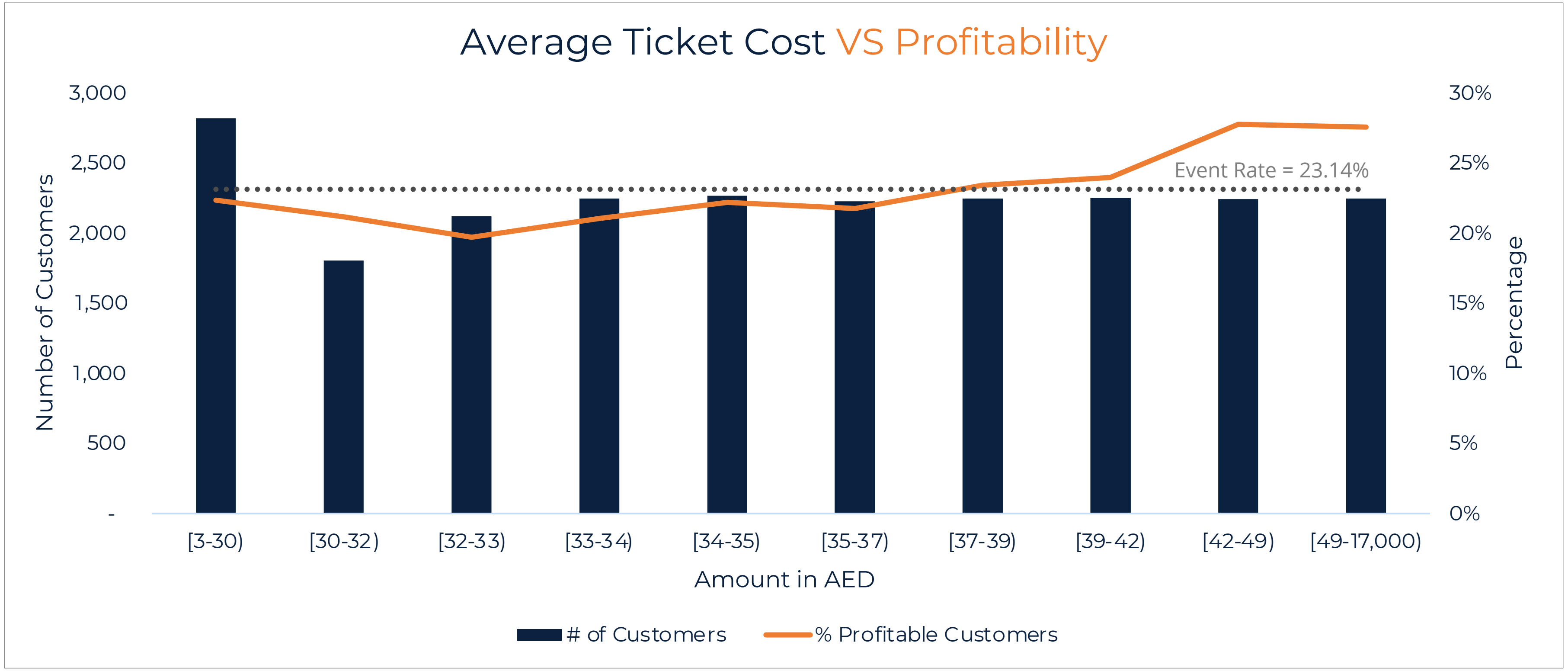
Bivariate Analysis

VOX Ticket Prices

Observations:

- ~45% of customers who watch movies in premium screens are profitable while only 23% of those who watch movies in regular screens are profitable.

AVERAGE TICKET COST AFFECTS PROFITABILITY



Bivariate Analysis

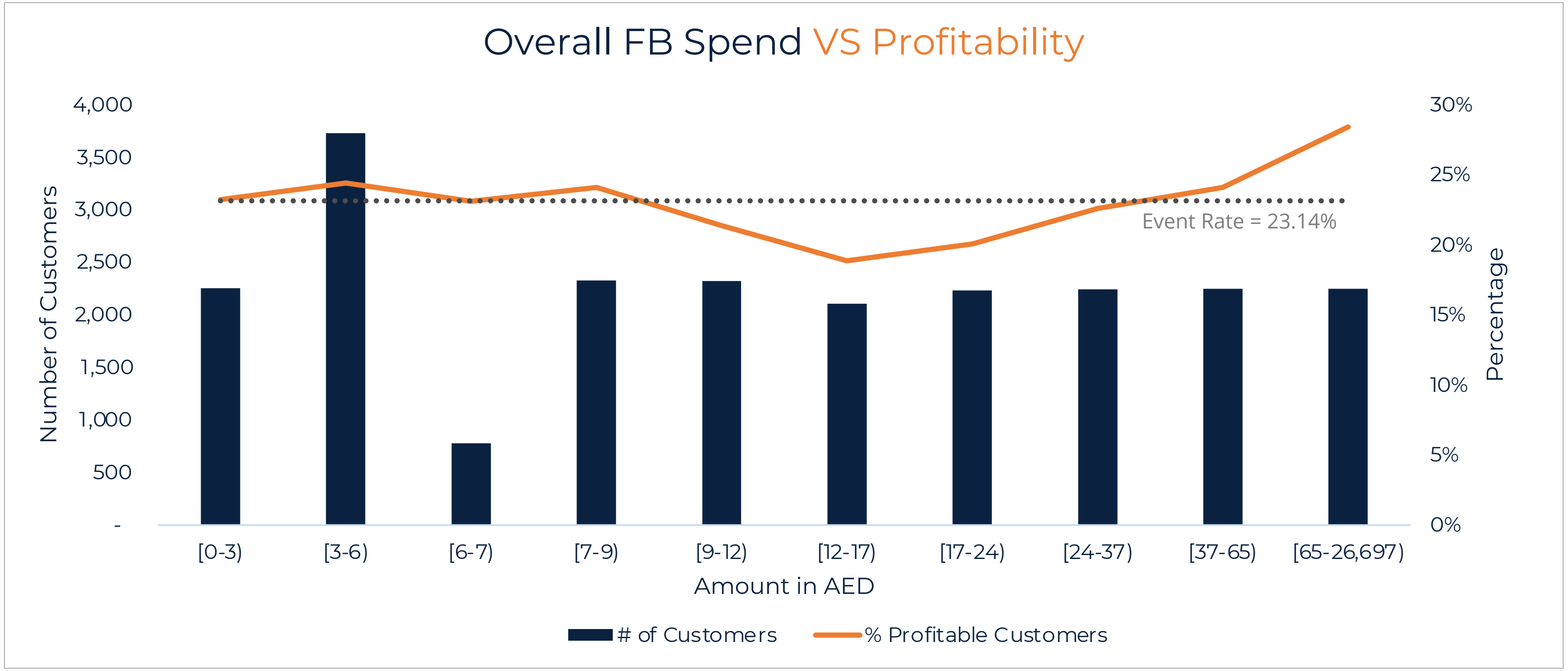
Observations:

- Customers spending AED 42- 49 per ticket on an average are highly profitable

VOX Ticket Prices

FOOD AND BEVERAGE SPEND AFFECTS PROFITABILITY

Bivariate Analysis

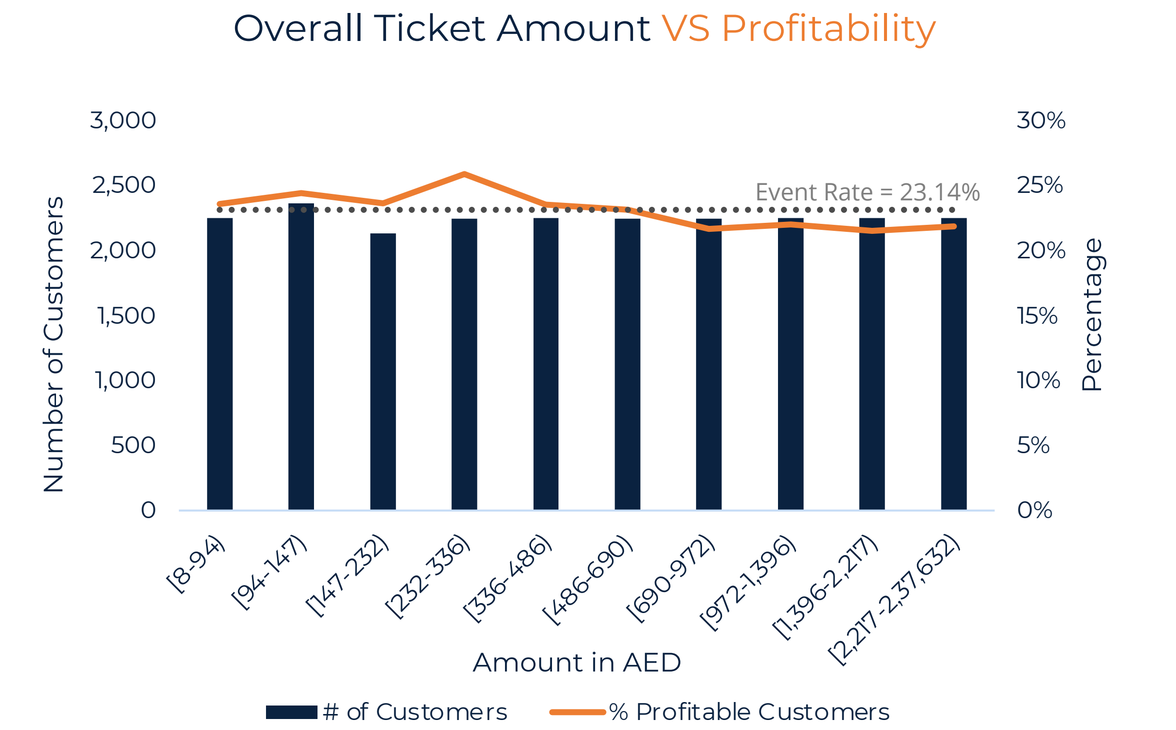


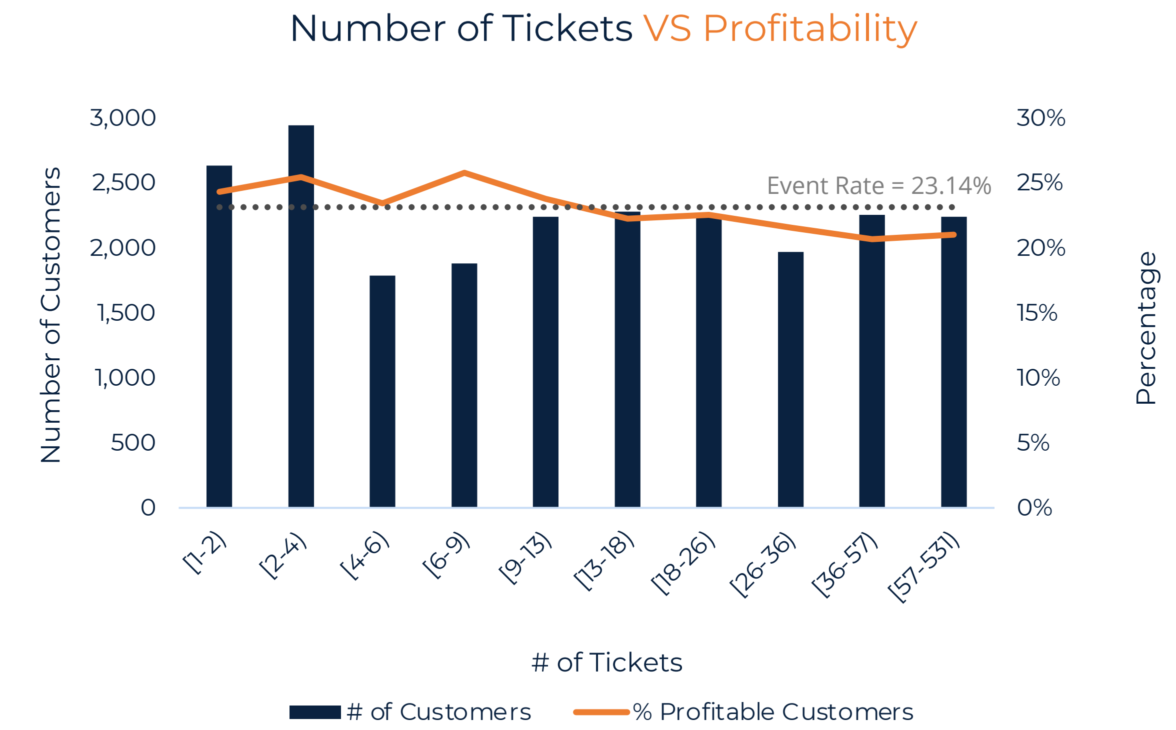
Observations:

- ~17% of the people spend AED 3-4 on Food & Beverage
- For customers spending AED 12 and above for F&B, we can see an increase in profitability when F&B spend increases

TICKETS BOUGHT AFFECT PROFITABILITY

Bivariate Analysis



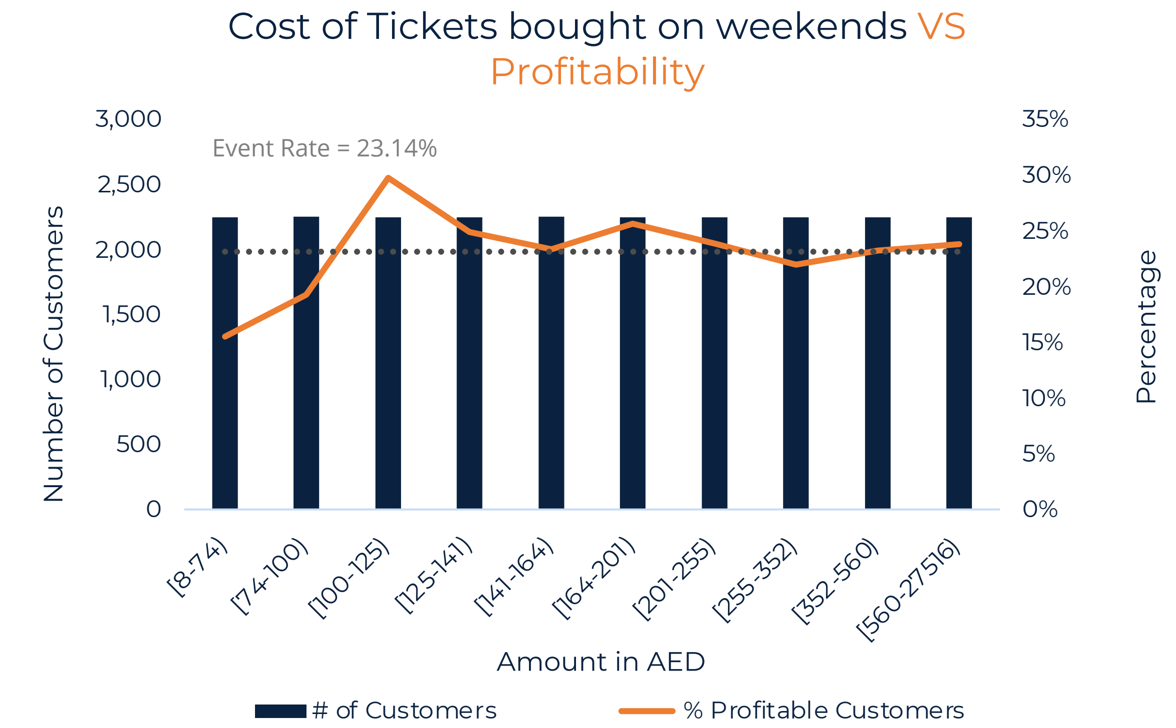


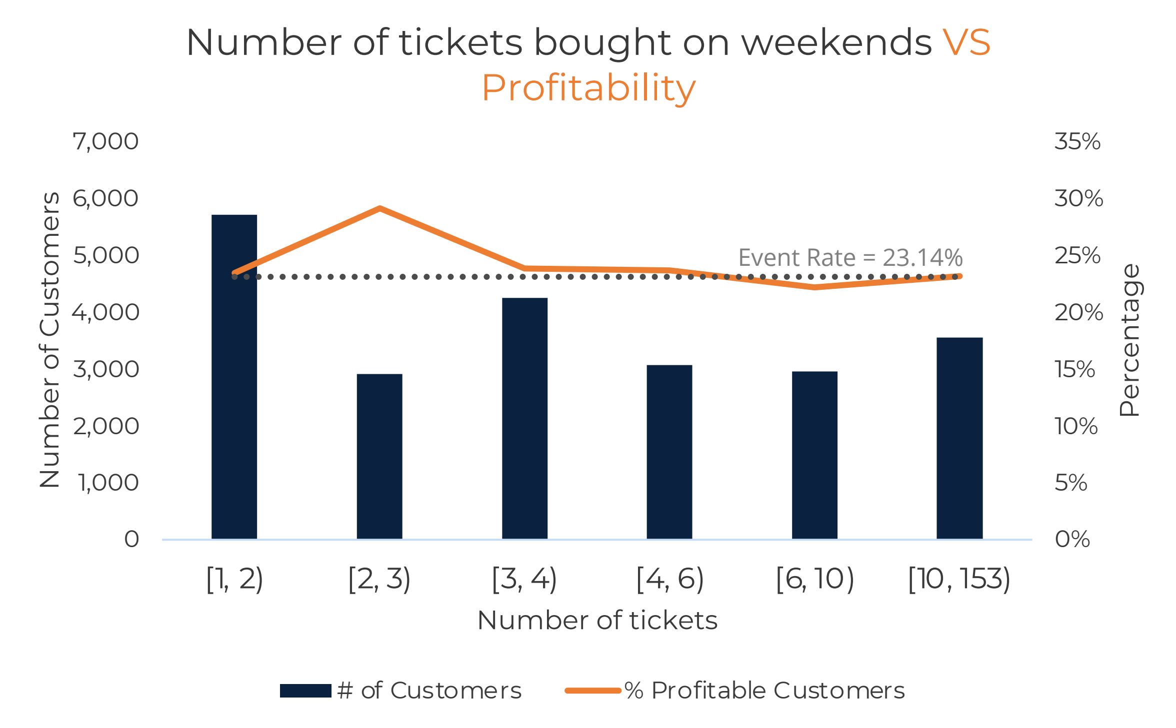
Observations:

1. People spending AED 232-336 have a high profitability index
2. ~25% of the customers buy 1-3 tickets, while only 9% buy 6-8 tickets yet are highly profitable as well

WEEKENDS AFFECT PROFITABILITY

Bivariate Analysis





Observations:

- Population spending about AED 100-125 for tickets on weekends have a higher chance of being profitable
- On weekends ~25% of customers just buy one ticket but people buying 2 tickets are the most profitable

CORRELATION ANALYSIS & VIF

USING VIF AND CORRELATION TO SELECT FEATURES

Removing data variables with a very high VIF >10 or correlation coefficient > 0.85 and iterating until we get satisfactory results

Correlation Matrix

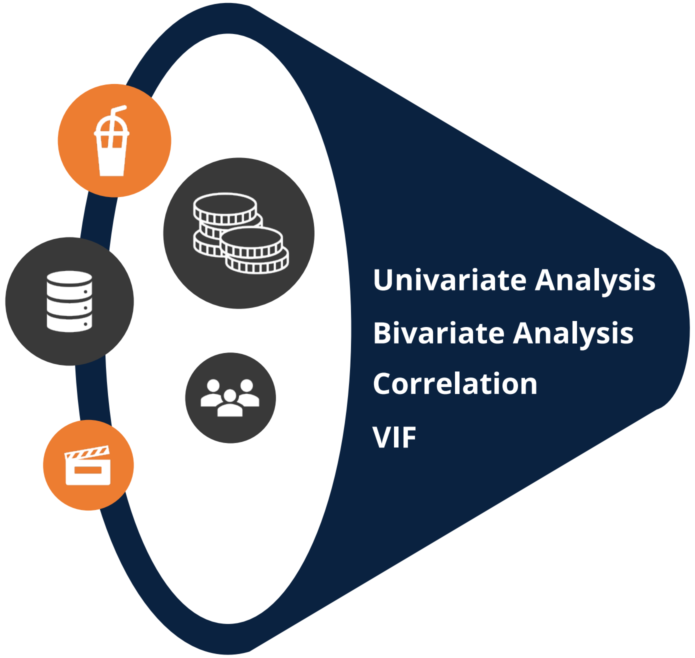
First Iteration

Sno	Features	VIF value
1	Last_60_days	22,687.12
2	Last_30_days	11,493.22
3	Last_90_days	11,222.60
4	Overall_Ticket_Amt	3,524.94
5	Booked_Amt	2,908.54
6	Overall_Spend	606.21
7	#Tickets	545.11
8	Pref_cinema_experience_#Ticket	454.55
9	Booked_Rdmption	111.43
10	Pref_movie_country_name_Spend	76.22
11	Pref_transaction_channel_Spend	46.64
12	Pref_transaction_channel_#Ticket	45.90
13	Pref_cinema_experience_Spend	43.99
14	#Movies_Watched	41.63
15	#Unique_Movies	41.21
16	Tickets_Weekend	40.80
17	Pref_movie_country_name_#Ticket	40.65

Sno	Features	VIF value
18	Pref_genre_name_Spend	27.75
19	#Weekends	24.75
20	Pref_film_rating_#Ticket	19.90
21	Pref_cinema_location_#Ticket	19.28
22	Pref_genre_name_#Ticket	18.08
23	Pref_cinema_location_Spend	14.281
24	Avg.Movie_Dur	8.87
25	Pref_film_rating_Spend	7.22
26	Avg_Tickt_Cost	5.79
27	Overall_FB_Spent	5.35
28	Is_internet_flag	3.63
29	Is_Action_flag	2.75
30	Is_mobile_flag	2.60
31	Is_Hollywood_flag	2.07
32	REVENUE_NAJM	1.70
33	New_Customer	1.52
34	Avg_Booking_Time	1.39

RELEVANT FEATURES WERE FILTERED OUT

53
Variables



→ **16**
Features

Final List of Features

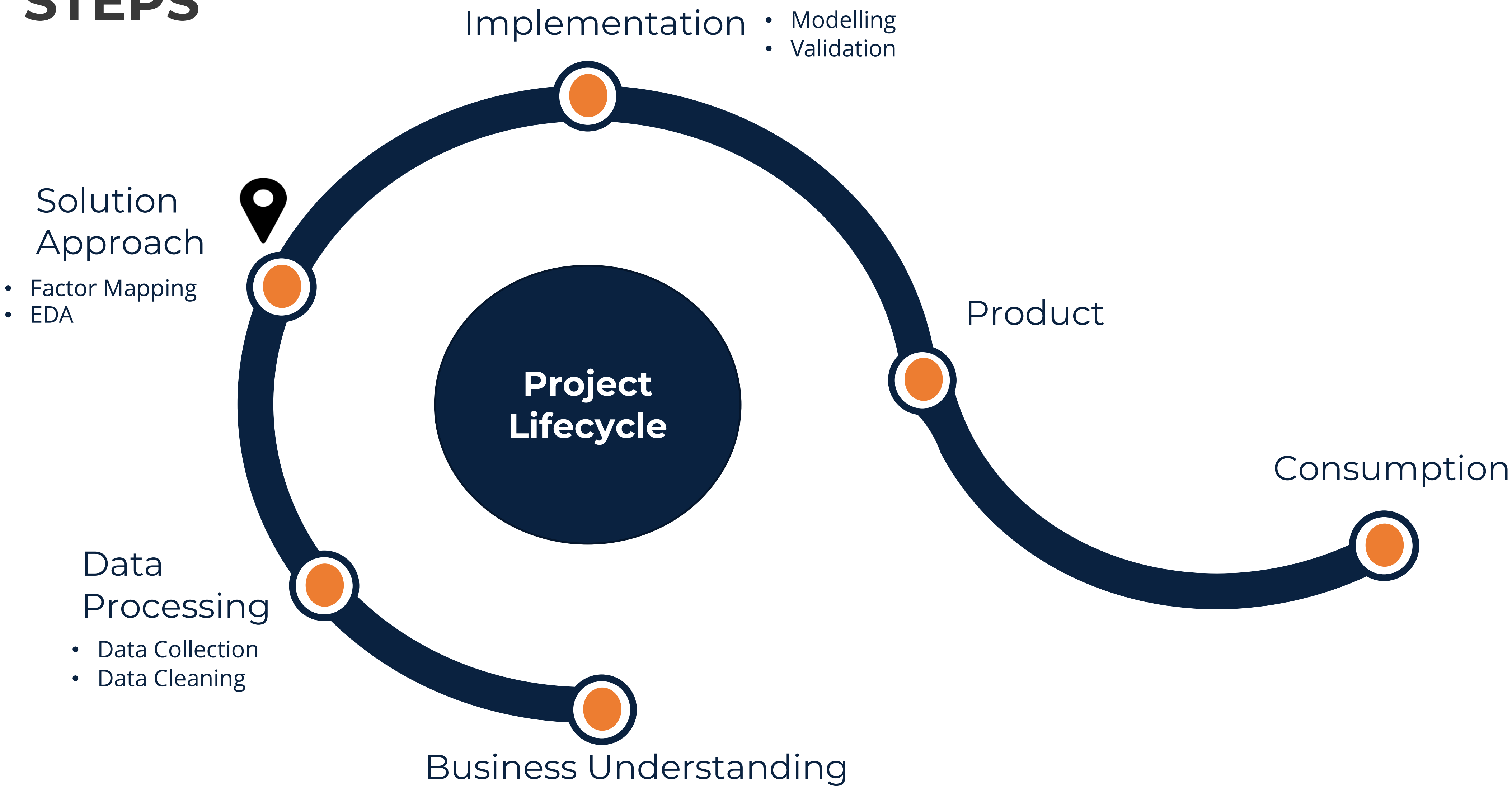
1. # of Tickets bought
2. # of Tickets bought on Weekends
3. Booking Amount
4. Booking Redemption Amount
5. Average Movie Duration
6. Average Ticket Cost
7. Transaction Channel (Internet Ticketing)
8. Transaction Channel (Mobile Phone)
9. Amount spent on preferred cinema location
10. Amount spent on preferred film rated movie
11. Amount spent on Food & Beverages
12. Watched an action movie or not
13. Watched a Hollywood movie or not
14. New Customer or not
15. # of Visits in Last 90 days
16. Average time taken to make a booking

16 features were selected from an exhaustive list of 53 variables through analysis

Proprietary and Confidential

NEXT STEPS

NEXT STEPS


Implementation
• Modelling
• Validation
Solution Approach
• Factor Mapping
• EDA
Project Lifecycle
Product
Consumption
Data Processing
• Data Collection
• Data Cleaning
Business Understanding

MODEL SELECTION

	K –Nearest Neighbours	Logistic Regression	Support Vector Machine	Decision Tree	Boosting Techniques	Random Forest
Outliers	SENSITIVE	SENSITIVE	ROBUST	ROBUST	SENSITIVE	ROBUST
Collinearity	SENSITIVE	SENSITIVE	SENSITIVE	ROBUST	ROBUST	ROBUST
Performance	LOW	LOW	MEDIUM	MEDIUM	HIGH	HIGH

- As our dataset contains a high number of features one decision tree cannot perform well and give the correct outcome
- It may memorise the training data in the decision tree if the parameters are not well tuned
- This can be overcome if we use Random Forest because it will build N number of decision trees and give the outcome based on polling
- Random forest and boosting is a combination of many decision trees thus, more compatible

THANK YOU

TheMathCompany

APPENDIX

DATA DICTIONARY (1/4)

S.No	Variable Name	Variable Type	Data Type	Variable Description
1	VOX_ID	Nominal	int64	Identification Number
2	Booked_Amt	Continuous	float64	Vox purchase amount
3	Booked_Rdmption	Continuous	object	Redeemed amount on the purchase
4	Avg_Booking_Time	Continuous	object	Average time period between ticket booking (in Days)
5	First_Transaction	Date Time	object	Customer's first transaction date (DD-MM-YYYY)
6	Last_Visit	Date Time	object	Customer's last visit to vox date (YYYYMMDD)
7	#Tickets	Discrete	object	Number of Tickets Purchased
8	#Movies_Watched	Discrete	object	Number of Movies watched
9	#Unique_Movies	Discrete	object	Number of Unique Movies Watched
10	Avg.Movie_Dur	Continuous	object	Average Movie Duration in Hrs
11	#Weekends	Discrete	float64	Number of tickets bought during Weekend
12	Cancl_Amt	Continuous	float64	Cancellation Amount
13	Cancl_Rdmption	Continuous	float64	Cancellation Redemption
14	Avg_Booking_Time_Cancl	Continuous	float64	Average Time period between ticket booking cancellation (in hours). Negative indicates that the person cancelled after the show started.
15	Cancl_Qty	Discrete	float64	Number of Tickets cancelled

DATA DICTIONARY (2/4)

S.No	Variable Name	Variable Type	Data Type	Variable Description
16	#Shows_Cancl	Discrete	float64	Shows Cancelled (Matinee, Morning, First Show, Second show)
17	#Cancl_Movies	Discrete	float64	Movies Cancelled
18	FB_Spend	Continuous	float64	Amount spent on Food and Beverages
19	FB_Rdmption	Continuous	float64	Amount Redeemed on Food & Beverages
20	transaction_channel	Nominal	object	Channel used to make the transaction
21	Is_internet_flag	Boolean	float64	Flag value to check if the transaction was made via internet ticketing (Yes - 1)
22	Is_mobile_flag	Boolean	float64	Flag value to check if the transaction was made via mobile phone (Yes - 1)
23	movie_country_name	Nominal	object	Cinema Industry Name
24	Is_Hollywood_flag	Boolean	float64	Flag value to check if it is a HOLLYWOOD Movie
25	genre_name	Nominal	object	Movie Genre
26	Is_Action_flag	Boolean	float64	Flag value to check if it is an ACTION Movie
27	film_rating	Nominal	object	Rating of the Movie
28	cinema_location	Nominal	object	Location of the Vox cinema theatre
29	cinema_experience	Nominal	object	Type of Cinema Experience

DATA DICTIONARY (3/4)

S.No	Variable Name	Variable Type	Data Type	Variable Description
30	Pref_transaction_channel_Spend	Continuous	float64	Total amount spend on making purchases using the preferred channel
31	Pref_transaction_channel_#Ticket	Discrete	float64	Number of tickets bought using the preffered channel
32	Pref_movie_country_name_Spend	Continuous	float64	Total Amount Spent in the Preferred Cinema Industry
33	Pref_movie_country_name_#Ticket	Discrete	float64	Number of tickets purchased in the Preferred Cinema industry
34	Pref_genre_name_Spend	Continuous	float64	Total amount spend on making purchases while visiting the preffered movie genre
35	Pref_genre_name_#Ticket	Discrete	float64	Number of tickets purchased for the preferred movie genre
36	Pref_film_rating_Spend	Continuous	float64	Total Amount Spent in the Preferred rating of film
37	Pref_film_rating_#Ticket	Discrete	float64	Number of tickets purchased in the Preferred rating of film
38	Pref_cinema_location_Spend	Continuous	float64	Total Amount Spent in the Preferred Cinema Location
39	Pref_cinema_location_#Ticket	Discrete	float64	Number of tickets purchased in the Preferred Cinema Location
40	Pref_cinema_experience_Spend	Continuous	float64	Total Amount Spent in the Preferred Cinema Experience
41	Pref_cinema_experience_#Ticket	Discrete	float64	Number of tickets purchased in the Preferred Cinema Experience

Proprietary and Confidential

DATA DICTIONARY (4/4)

S.No	Variable Name	Variable Type	Data Type	Variable Description
42	REVENUE_NAJM	Continuous	float64	Revenue generated on that customer ID
43	Overall_Ticket_Amt	Continuous	object	Total cost of tickets
44	Overall_Tickt_Cncld_Amt	Continuous	float64	Total ticket cancelled amount
45	Avg_Tickt_Cost	Continuous	float64	Average ticket cost
46	Overall_FB_Spent	Continuous	float64	Total amount spent on Food and Beverages
47	Tickets_Weekend	Continuous	object	Amount spent on Tickets during weekends
48	Overall_Spend	Continuous	float64	Total amount spent
49	New_Customer	Boolean	float64	Whether the customer is new or not
50	Avg_Cost_per_Ticket_Cancld	Continuous	float64	Average cost of cancellation per ticket
51	Last_30_days	Discrete	float64	# of visits in last 30 days
52	Last_60_days	Discrete	float64	# of visits in last 60 days
53	Last_90_days	Discrete	float64	# of visits in last 90 days

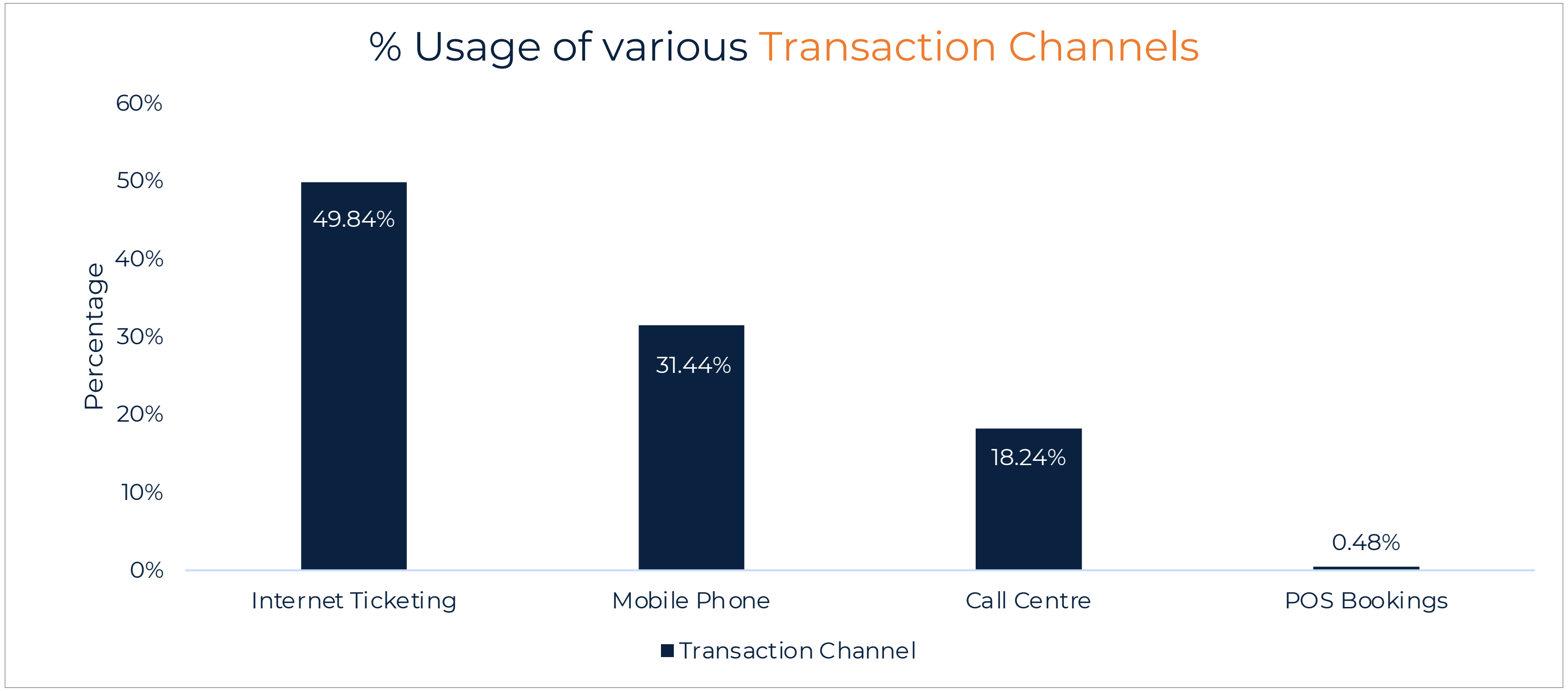
DATASET - DETAILED SUMMARY

All data excluding profitability column

	Total	Count		% missing values	of variables	Total
Numerical	39	23	Continuous	>90%	6	23
				>1% and <=90%	6	
				<=1%	11	
		16	Discrete	>90%	3	16
				>1% and <=90%	5	
				<=1%	8	
Categorical	12	7	Nominal	>90%	0	7
				>1% and <=90%	2	
				<=1%	5	
		5	Boolean	>90%	1	5
				>1% and <=90%	4	
				<=1%	0	
Time - Series	2	2	DateTime	>90%	0	2
				>1% and <=90%	0	
				<=1%	2	
	53				53	

>90% missing Values	10
< 5% missing Values	26
>=5% and <=90% missing va	17
	53

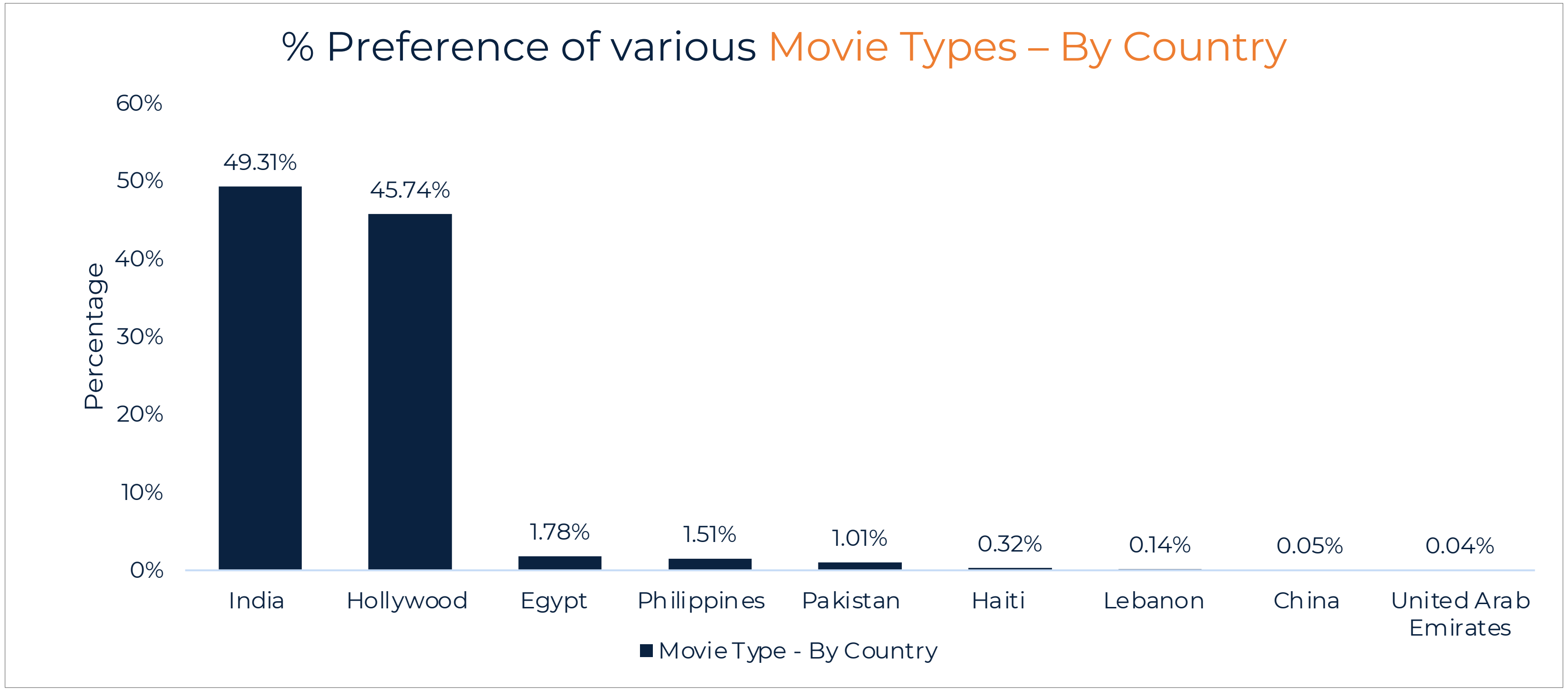
TICKETS BOUGHT VIA INTERNET TICKETING IS HIGH



Observations:

~50% of the tickets are booked via Internet ticketing followed by ~31% of all tickets booked via mobile phone

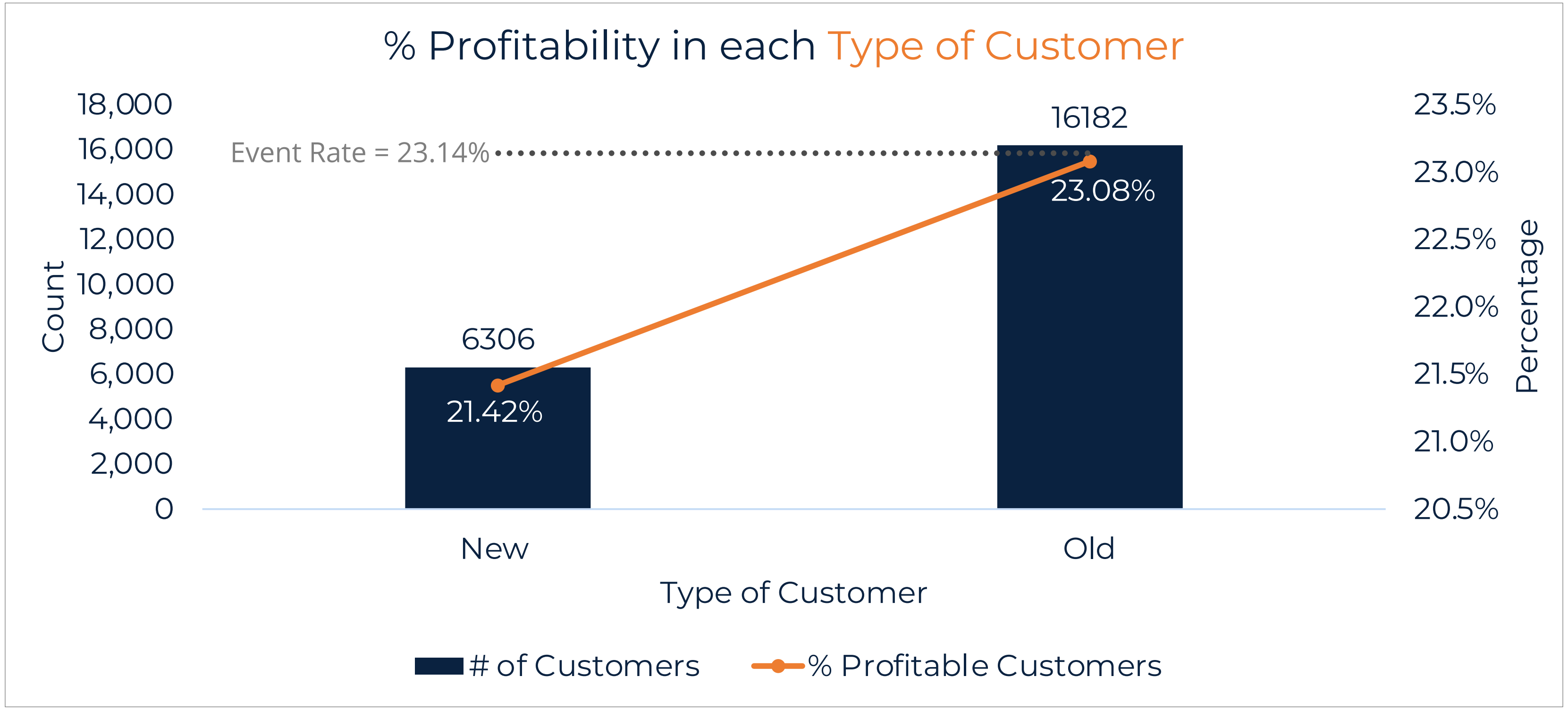
INDIAN MOVIES ARE THE MOST PREFERRED



Observations:

~50% of the customers prefer watching Indian movies followed by ~46% of them preferring Hollywood movies to watch

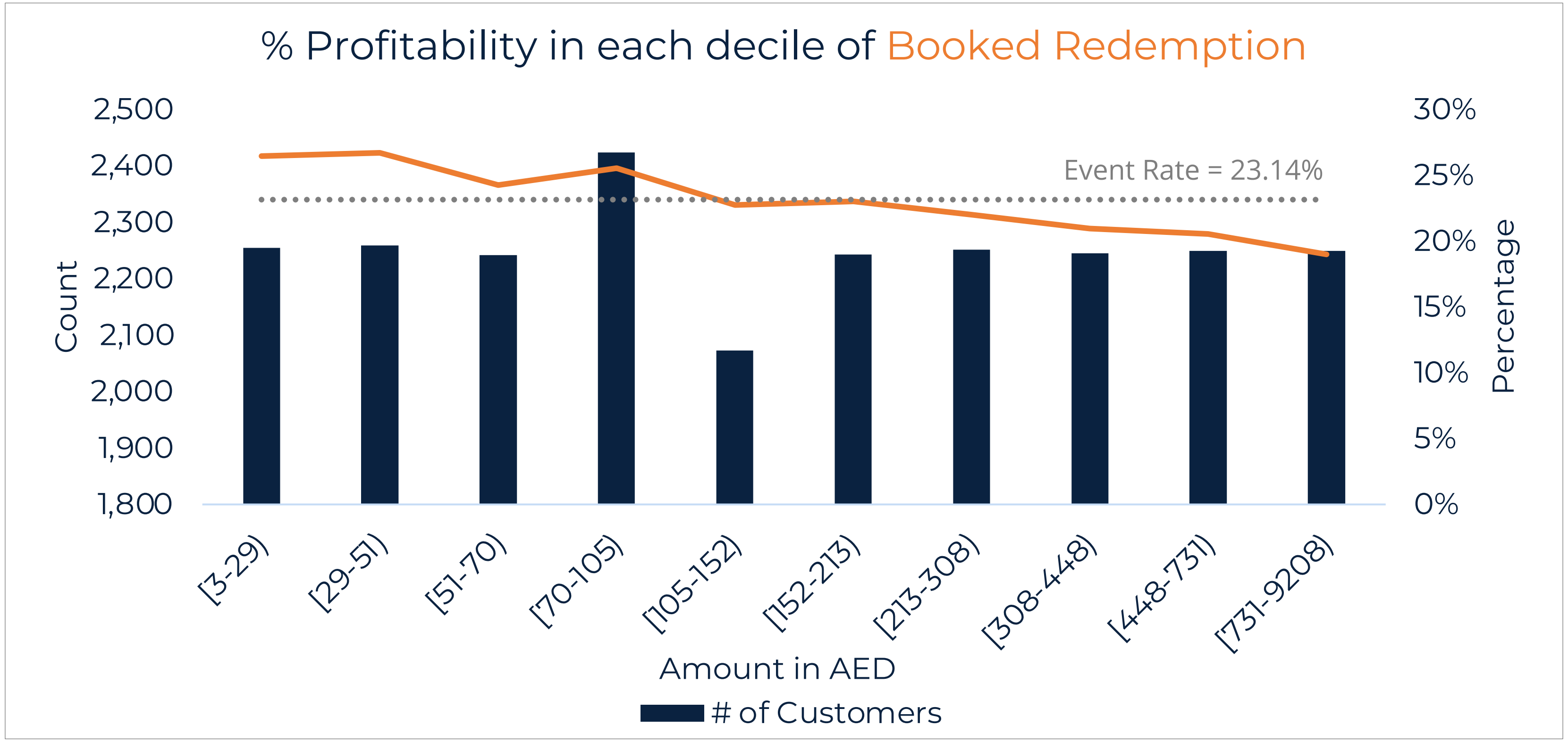
NEW CUSTOMER NOT AFFECTS PROFITABILITY



Observations:

~23% of the already existing customers who are profitable while only 21% of the new customers are profitable.

CUSTOMERS WHO AVAIL OFFERS ARE NOT PROFITABLE

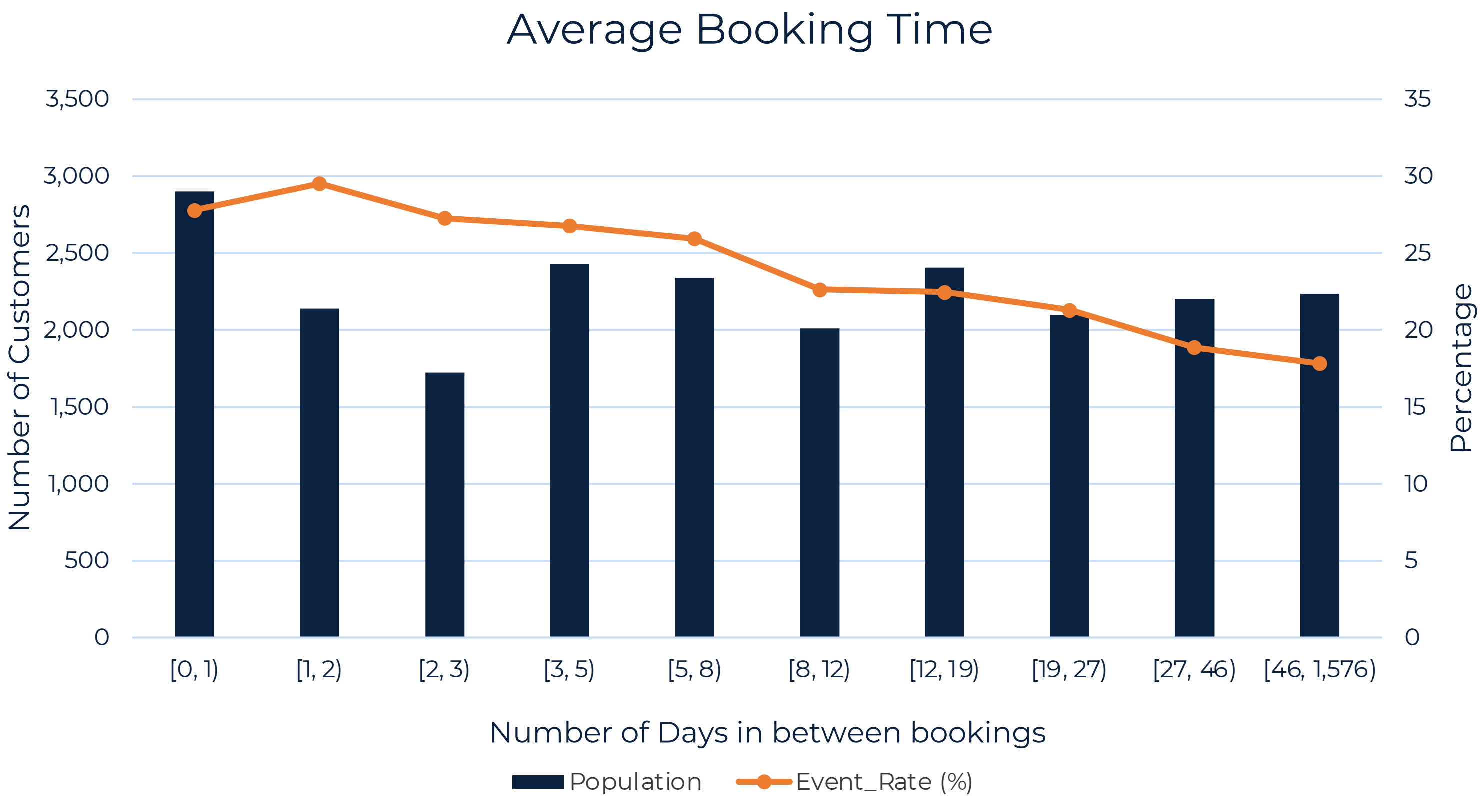


Observations:

~ 26% of people who redeem around AED 70-104 on their booking amount are profitable

After AED 152, whoever redeemed on their bookings have a downtrend in profitability

CUSTOMERS VISITING FREQUENTLY ARE PROFITABLE



Observations:

Higher the average number of days in between bookings, lower is the profitability.

IMPUTATION

Description of Widgets used in Data Imputation

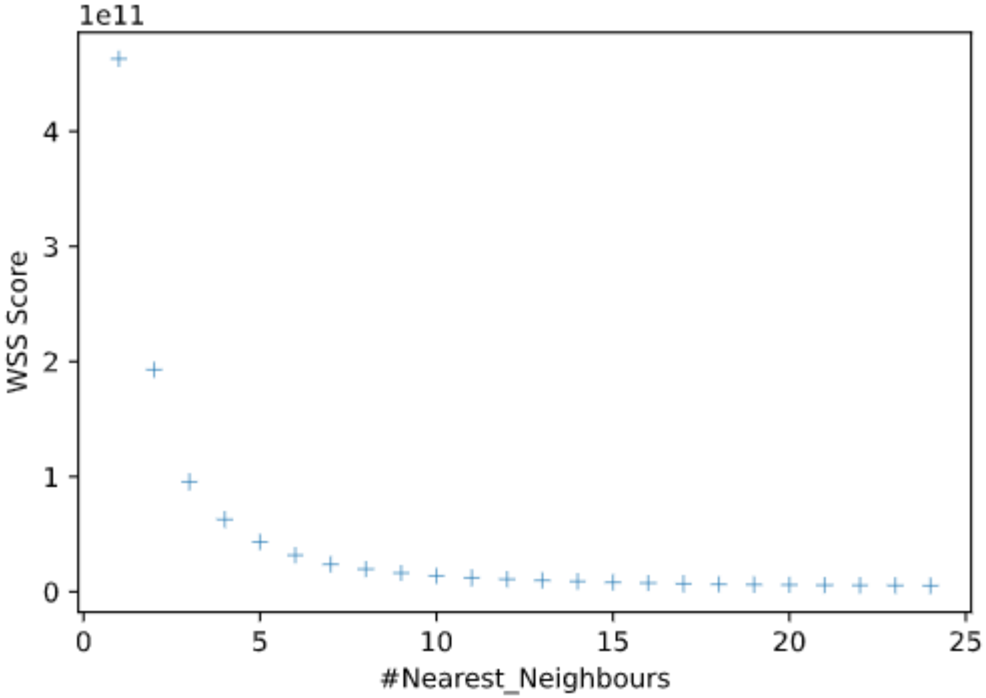
- **Missing Values per Column:** Returns a table consisting data of Missing Count and Missing Percent for each column.
- **Impute Categorical:** Imputes missing values in categorical columns with mode.
- **Impute Boolean:** Imputes missing values in Boolean columns according to the data present in categorical columns.
- **Impute Continuous:** Imputes missing values in continuous columns whose missing data percent is ≤1% with median.
- **KNN_Impute:** Imputes missing values in continuous columns whose missing data percent is >1% with KNN algorithm.
- **Assembling_Columns:** Concatenates categorical, boolean and continuous columns.
- **Datatype_Conversion:** Converts columns datatype to the required datatype (int, float etc).

KNN IMPUTATION

Steps in KNN

- If the percentage of missing data in a column is greater than 1%, we imputed missing data with KNN imputer
- To know the k-value, we plotted an elbow curve
- We chose 5 nearest neighbours and imputed missing values with KNN

Elbow Plot



SELECTION OF RELEVANT FEATURES - OBSERVATIONS

9

Iterations of VIF

Overall Spend = Overall Ticket Amount + Overall FB Spend

Overall Ticket Amount → Booked Amount + Booked Redemption

Dropped Columns

6

Features highly correlated with Number of Tickets (Corr. Coeff. > 0.85)

4

Features highly correlated with Booking Amount (Corr. Coeff. > 0.88)

Last 30 days

Last 60 days

Highly correlated

Highly correlated

Last 90 days

- 12 features were dropped as they were highly correlated with other features
- Overall Ticket Amount and Overall spend can be deduced from other columns hence it is dropped

CORRELATION ANALYSIS TO FEATURE SELECTION

Highly Correlated variables (>0.85)

Booked Amount

- Pref_transaction_channel_Spend
- Pref_movie_country_name_Spend
- Pref_cinema_experience_Spend
- Pref_genre_name_Spend

 # Tickets

- Pref_transaction_channel_#Ticket
- Pref_movie_country_name_#Ticket
- Pref_film_rating_#Ticket
- Pref_cinema_location_#Ticket
- Pref_cinema_experience_#Ticket
- Pref_genre_name_Spend_#Ticket

If correlation coefficient > 0.9 then there is multicollinearity

*Multicollinearity - multiple factors that are correlated to each other

Proprietary and Confidential

OUTLIERS – DETAILED VIEW

S.no	Data Variables	# of Outliers
1	#Movies_Watched	1813
2	Avg_Booking_Time	1796
3	Pref_cinema_location_Spend	1787
4	Pref_cinema_location_#Ticket	1713
5	Pref_transaction_channel_#Ticket	1688
6	Pref_genre_name_Spend	1679
7	Booked_Amt	1677
8	Overall_Ticket_Amt	1671
9	Overall_Spend	1667
10	Pref_transaction_channel_Spend	1658
11	Pref_genre_name_#Ticket	1658
12	Pref_cinema_experience_Spend	1653
13	Pref_movie_country_name_Spend	1631
14	Pref_film_rating_Spend	1628
15	#Tickets	1621
16	Pref_film_rating_#Ticket	1598
17	Pref_cinema_experience_#Ticket	1596
18	Booked_Rdmption	1563
19	Avg_Tickt_Cost	1548
20	Pref_movie_country_name_#Ticket	1535
21	#Unique_Movies	1459
22	Last_90_days	1421
23	Last_60_days	1419
24	Last_30_days	1418
25	REVENUE_NAJM	1380
26	Overall_FB_Spent	1068
27	#Weekends	1006
28	Tickets_Weekend	989

ARCHITECTURE DIAGRAM

Data Dictionary

Bivariate Analysis

Correlation Matrix

VOX Ticket Prices

Dataset - After Imputation

Factor Mapping

Univariate Analysis

Data Dictionary - After Cleaning

Data

Cleaned Data