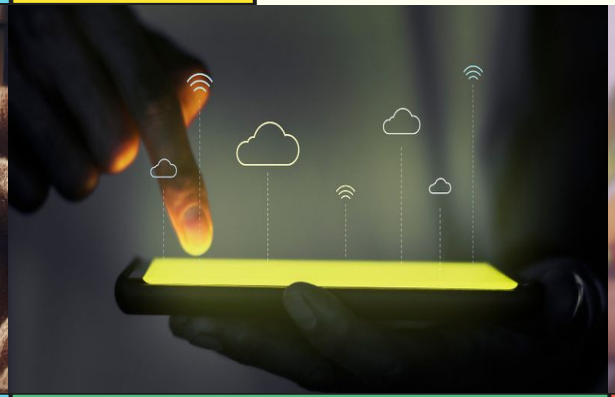
By: Yisi Lu, Lim Li, Kenny Lei, Alex Fan

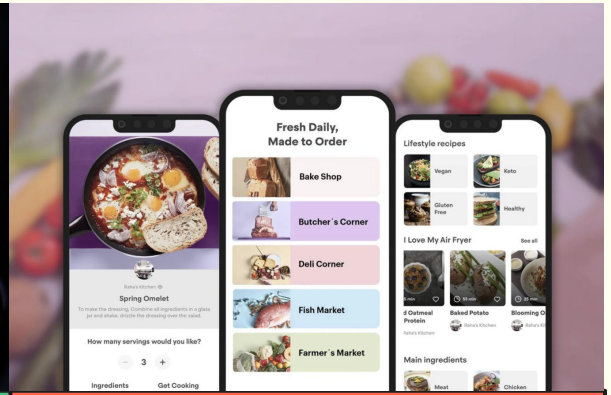
Kaggle Project

Datasets

Dataset 1: Coffee Revenue



Dataset 2: Advertisement



Dataset 3: Food App Business

About the Data: Coffee Shop

Overview:

This dataset contains 2,000 rows of data from coffee shops, offering detailed insights into factors that influence daily revenue.

Data Structure:

The dataset consists of only one table so there is only one primary key. The primary key for our dataset is the day number. Each day in the table is unique. Our table consists of numeric data types.

Business Problems:

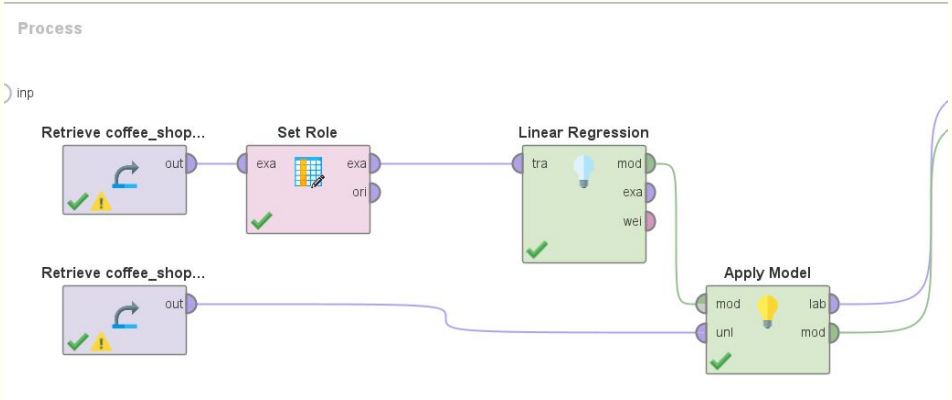
- Figure out which factors contribute most to maximizing revenue
- Figure out where to potentially cut costs
- Where to make improvements

Data Trends:

- Higher revenue days tend to have either a higher customer count, high order value, or a combination of both
- Higher revenue days generally have higher marketing spend but a higher marketing spend does not guarantee higher revenue

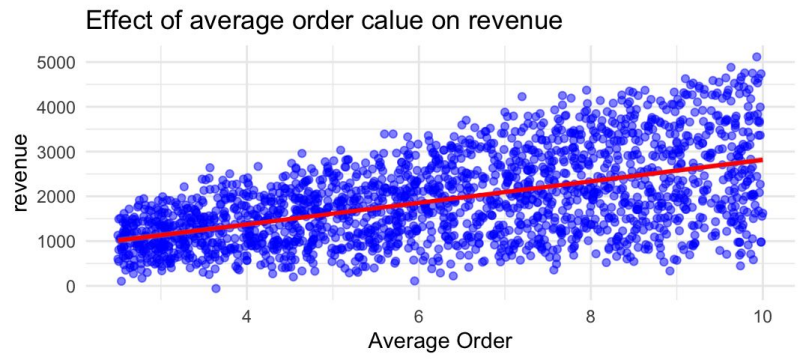
Rapidminer Analysis: Linear Regression

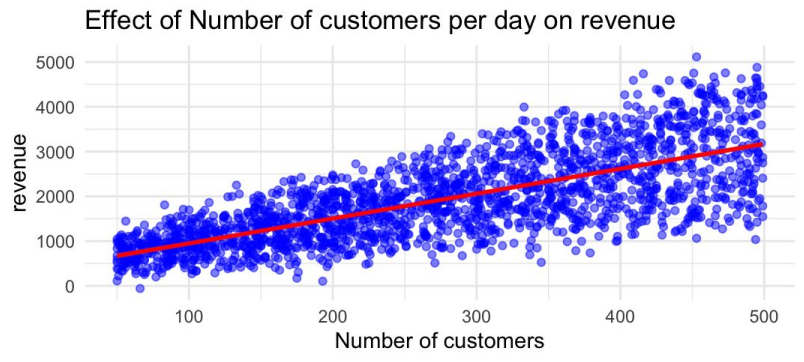
Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
Number_of_C...	5.564	0.066	0.742	1.000	84.085	0	****
Average_Orde...	242.026	3.948	0.541	1.000	61.310	0	****
Marketing_Sp...	1.453	0.060	0.213	0.999	24.142	0	****
Location_Foot...	0.033	0.031	0.009	1.000	1.043	0.297	
(Intercept)	-1504.680	39.000	?	?	-38.582	0	****



- Split Data into 70% training 30% testing
- Average Order Value, Marketing Spending, and Number of Customers are significant
- Average Order Value has the highest coefficient of 242.026
- Number of employees, foot traffic, and hours of operation are not significant

Graphs:


Effect of average order calue on revenue
revenue
5000
4000
3000
2000
1000
0
4
6
8
10
Average Order


Effect of Number of customers per day on revenue
revenue
5000
4000
3000
2000
1000
0
100
200
300
400
500
Number of customers


Effect of Marketing spend on revenue
revenue
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
Marketing_Spend_Per_Day

About the Data: Advertising

Overview:

This dataset contains 1,000 rows of data from an online advertising campaign, providing detailed insights into factors that influence user engagement and ad click-through rates.

Data Structure:

The dataset consists of one table. To ensure data uniqueness, we use Ad Topic Line and Timestamp as a composite primary key. While the Timestamp alone may be unique in this table, it might not be unique in a larger dataset or across multiple advertising campaign records. Therefore, combining these two fields ensures uniqueness.

Business Problems:

- Identify key factors that contribute to higher ad click-through rates

Data Trends:

- Lower-income users have higher ad click rates, while higher-income users engage less.
- Users with shorter online time are more likely to click ads, while long-time users have the lowest engagement.

SQL and R Analysis

1. R was used to find which categories have a logistic regression with clicked on ads.

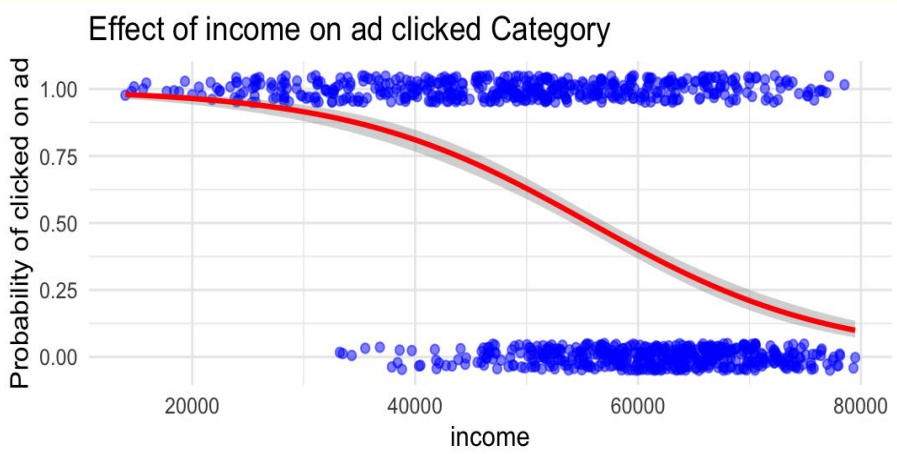
2. In R, three datasets related to "Clicked in Ad" were identified: "Daily Time Spent on Site," "Age," and "Area Income."

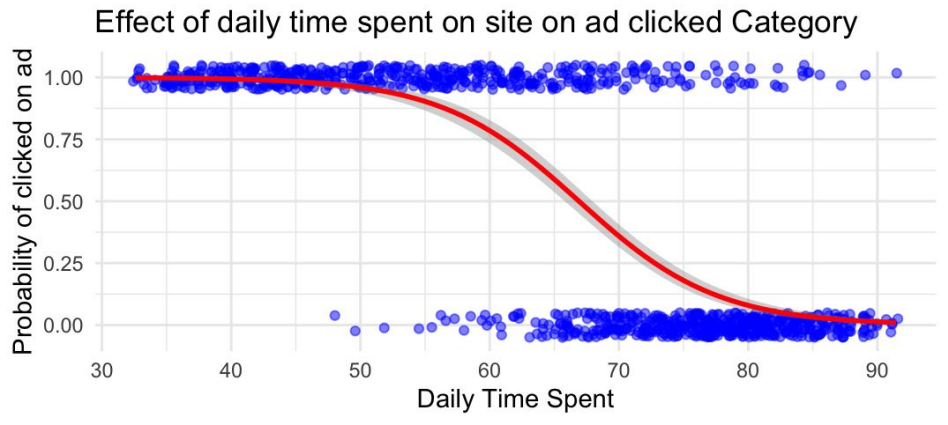
3. In SQL, further analysis was conducted on "Clicked Rate" and "Income," as well as "Time Spent," "User Count," "Ad Clicks," and "Ad Click Rate."

Graphs:

	Age	Income_Group	Click_Rate
1	26	High Income	57.1428571428571
2	39	Low Income	100.0
3	48	Middle Income	92.6470588235294
4	35	Most High Income	27.6190476190476

	Time_Spent_Group	User_Count	Ad_Clicks	Ad_Click_Rate
1	Long Time Spent	462	55	11.9
2	Moderate Time Spent	305	214	70.16
3	Short Time Spent	233	231	99.14





About the Data: Food App Business

Overview:

This dataset contains 2,205 records from a food app, showing detailed insights into factors which have an impact on overall profitability..

Data Structure:

The dataset contains only one table, with 27 attributes originally. The uniqueness of each attribute was tested by the SELECT DISTINCT command, and a unique identifier was added to each observation for future investigation. There are only numeric values in this table.

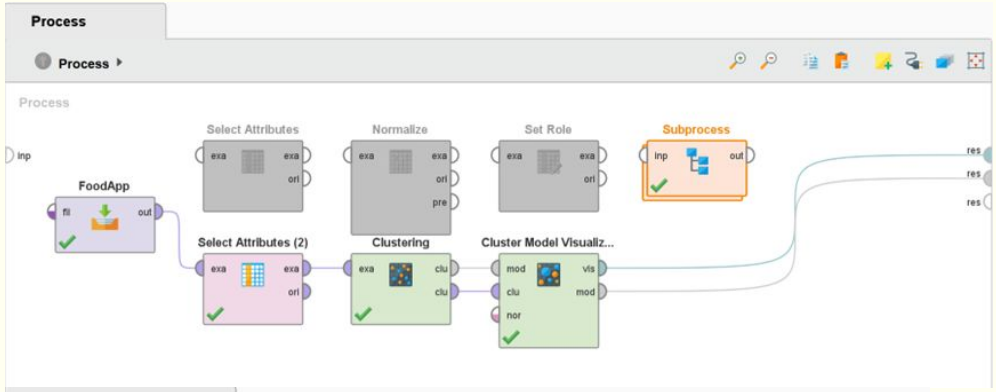
Business Problems:

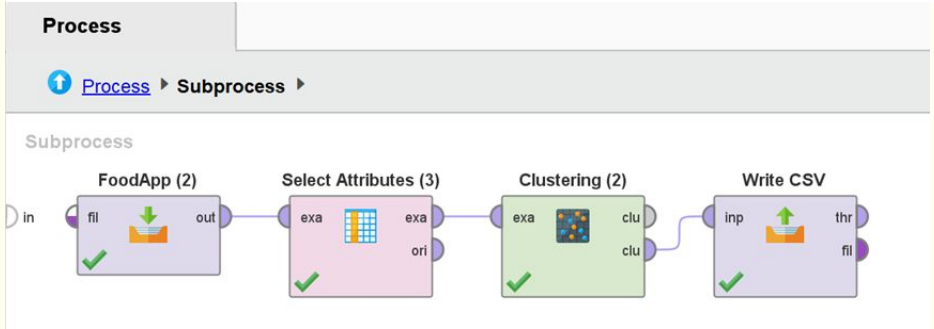
- Figure out customer behavior and spending patterns
- Identify relationships between factors
- Make new strategies and improve customer retention

Data Trends:

- Customers accept more campaigns and spend more on discounts suggests that price-sensitive customers are more likely to be influenced by promotions.
- Higher-income clusters tend to make more web purchases than in-store orders. They have low reliance on discounts, while showing high engagement with marketing campaigns.

RapidMiner and SQL Analysis






```sql
SELECT
    cluster,
    SUM(TotalNoOfCampaignAccepted) AS [total campaigns accepted],
    SUM(NoOfDealsWithDiscount) AS [total spending on discounts]
FROM
    [FoodAppBusiness-clustered]
GROUP BY
    cluster;
```

	cluster	total campaigns accepted	total spending on discounts
1	cluster_0	140	792
2	cluster_1	44	948
3	cluster_2	90	1303
4	cluster_3	20	574
5	cluster_4	63	1182
6	cluster_5	303	313

- Check mean value and standard deviation to ensure there are no outliers
- Exclude the attributes that are less significant
- Apply k-means clustering and divide the dataset into 6 groups with different features
- Focus on the influence of campaigns

Graphs

Cluster Model (Clustering) | ClusterModelVisualizerIOObject (Cluster Model Visualizer)

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
MonthlyIncome	83821.795	33170.070	58245.188	45580.141	70302.327	18939.306
Age	52.080	48.395	54.708	52.786	52.768	44.049
NoOfChildren	0.028	0.561	0.195	0.343	0.087	0.502
NoOfTeenager	0.111	0.276	0.589	0.457	0.347	0.094
AmountSpendOnWines	681.538	39.753	408.150	157.849	579.753	10.989
AmountSpentOnFruits	65.747	5.800	27.602	9.114	51.187	5.849
AmountSpentOnMeat	489.979	26.962	138.155	55.005	321.187	20.898
AmountSpentOnFish	94.132	9.796	34.424	14.545	75.182	7.868
AmountSpentOnSweet	72.618	5.825	25.708	9.406	51.524	6.226
NoOfDealsWithDiscount	1.087	2.126	3.307	2.876	1.975	2.166
NoOfWebPurchase	5.271	2.188	5.619	3.779	5.623	1.985
NoOfStorePurchase	8.455	3.260	7.452	4.579	8.506	2.728
NoOfWebVisitsMonth	2.538	6.704	5.264	6.221	3.716	7.268
TotalNoOfCampaignAccepted	1.052	0.099	0.228	0.153	0.349	0.075