Final Project Whitepaper
Analyses of 3 Kaggle Datasets

Section 1: Executive Summary
Data Set 1

The first dataset contains data from a coffee shop consisting of 2000 days of service. Each day includes insights into factors that impact daily revenue. The data consists of one table with one primary key and numeric values. This data was interesting and could be useful to identify what attributes lead to the greatest revenue. Number of customers, average order value, and daily marketing spending are the biggest contributing factors in increasing revenue. Hours of operation, number of employees, and location foot traffic have little to no impact on the revenue. Marketing spending generally leads to higher revenues and more customers but it is not guaranteed. This information can be applied by trying to find ways to increase average order value and potentially cut costs by lower hours of operation and employee size. Based on the data, marketing campaigns have room for improvement and the spending can be more optimally used as there is no direct correlation between marketing spending and revenue.

Data Set 2

This dataset contains users' personal information, browsing behavior, and ad click activity. By analyzing this data, we can understand user ad-clicking behavior patterns, which helps improve the accuracy and conversion rate of ad placements.

Data Set 3

This dataset is about the operations of an online food app business, providing values of factors that affect consumer behaviour, which we are interested in and investigating in this paper. The dataset has only one table, and it contains only numeric values representing customer demographics, most consumed food, and engagement with promotions. This dataset is applicable in seeking trends in customers' preference and helps us identify what drives higher spending and customer retention. Monthly income, number of children, and age are significant factors that has great impact on consumer spending. Higher-income individuals allocate more funds to premium food categories such as meat, wine, and seafood. Web purchases and store purchases vary across clusters, with some groups preferring online transactions while others rely on in-store purchase. By exploring the influence of discount deals and marketing campaigns on customer engagement, the total spending varies across different segments. This suggests that these two attributes are less significant than the characteristics of demographics. BY the end of analysis, we get the picture that targeted marketing strategies and personalized promotions can enhance customer retention and increase revenue. Businesses can optimize customers' spending by identifying high-value customers and tailoring promotions to maximize their engagement.

Section 2: Analysis of 3 Data Sets

Data Set 1: [Coffee Shop 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.

coffee_shop_revenue		CREATE TABLE "coffee_shop_revenue" ("Days" INTEGER, "Number_of_
Days	INTEGER	"Days" INTEGER
Number_of_Customers_Per_Day	INTEGER	"Number_of_Customers_Per_Day" INTEGER
Average_Order_Value	INTEGER	"Average_Order_Value" INTEGER
Operating_Hours_Per_Day	INTEGER	"Operating_Hours_Per_Day" INTEGER
Number_of_Employees	INTEGER	"Number_of_Employees" INTEGER
Marketing_Spend_Per_Day	REAL	"Marketing_Spend_Per_Day" REAL
Location_Foot_Traffic	INTEGER	"Location_Foot_Traffic" INTEGER
Daily_Revenue	REAL	"Daily_Revenue" REAL

Analysis with Rapid Miner and SQL:

Rapid Miner was used to run a linear regression on the data to predict daily revenues based on these parameters. We split the coffee shop revenue data into 70% training data and 30% scoring data. The analysis revealed that the biggest contributor to daily revenue was the number of customers, average order value, and marketing spend per day. Location foot traffic, hours of operations, and number of employees did not have a significant impact on daily revenue. Though number of customers, average order, and marketing spend are all significant, average order value has the highest coefficient of 242.026 which means for every extra dollar spent in average order value the daily revenue goes up by 242.026. This means that average order value is the greatest contributor out of the significant attributes.

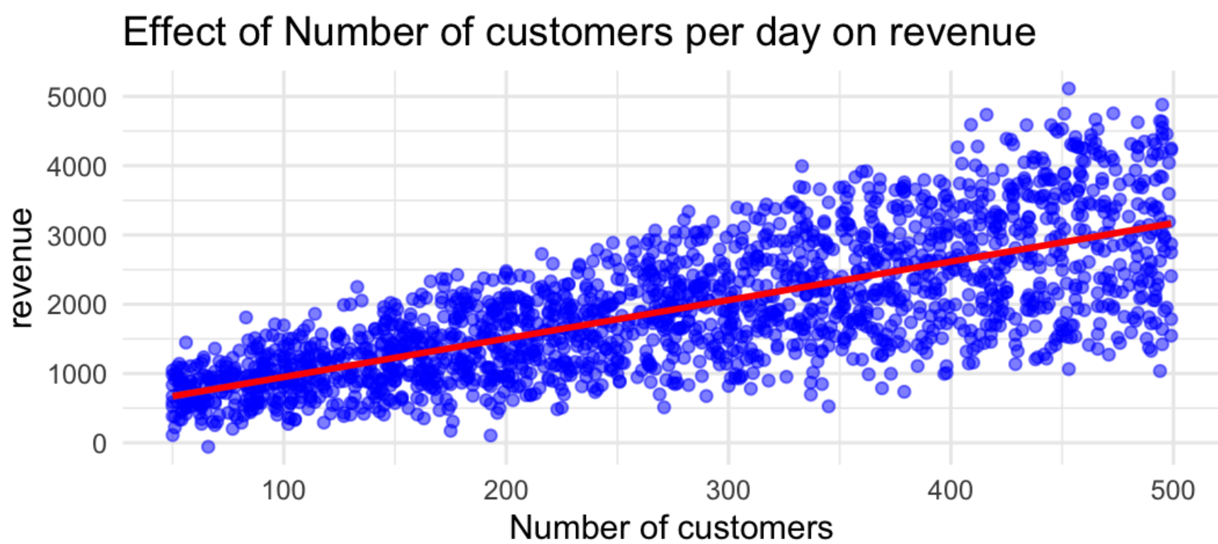
Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
Number_of_C...	5.564	0.066	0.742	1.000	84.085	0	****
Average_Orde...	242.026	3.948	0.541	1.000	61.310	0	****
Marketing_Sp...	1.453	0.060	0.213	0.999	24.142	0	****
Location_Foot...	0.033	0.031	0.009	1.000	1.043	0.297	
(Intercept)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

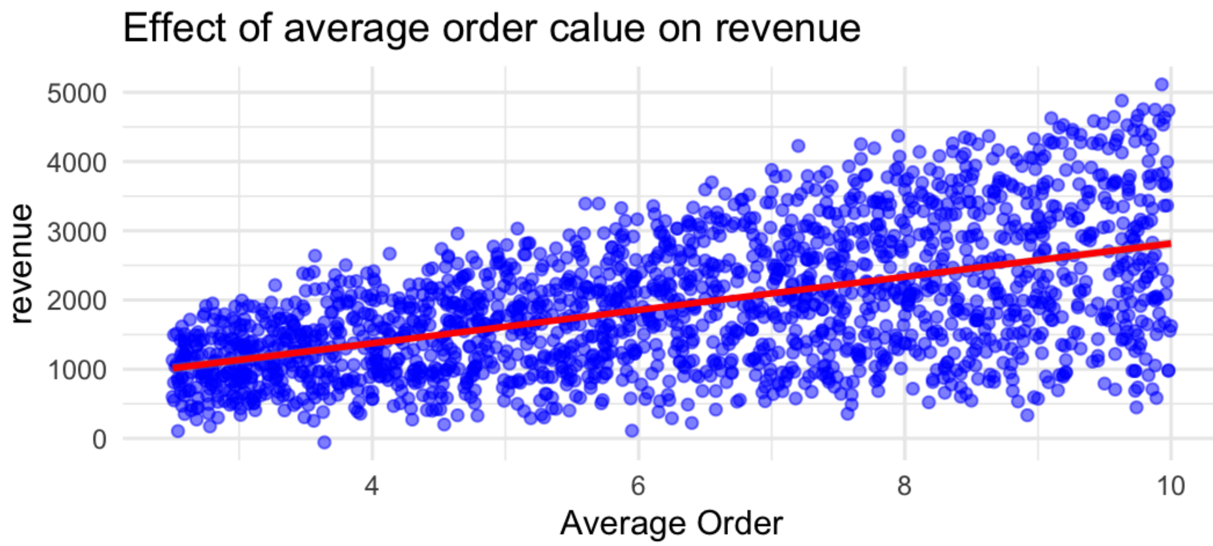
Data Trends:

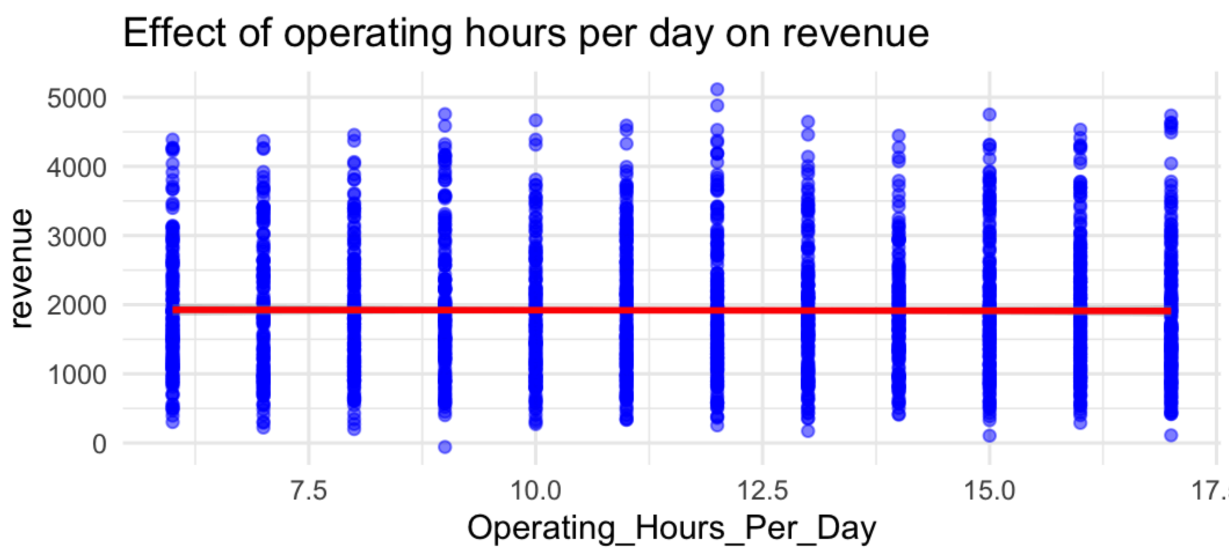
When analyzing the data's most common and grouped information, we can see that the highest 3 days of revenue were $5114.6 on day 1594, $4881 on day 1720, and $4756.55 on day 1144. Even though marketing spend per day is said to have a significant effect, it does not have a direct effect on the number of customers per day. Even though usually a higher marketing spend day means more revenue, there is not a direct positive correlation between marketing spend and daily customers. Location foot traffic, employee number, and operating hours had little to no effect on the amount of customers. For the three highest revenue days, the average order value is on the higher end in the $9 range and the customer number is in the 400 range. The marketing spend is also in the upper range. Based on these data trends, it seems that a combination of high average order value and number of customers is what generates the most revenue.

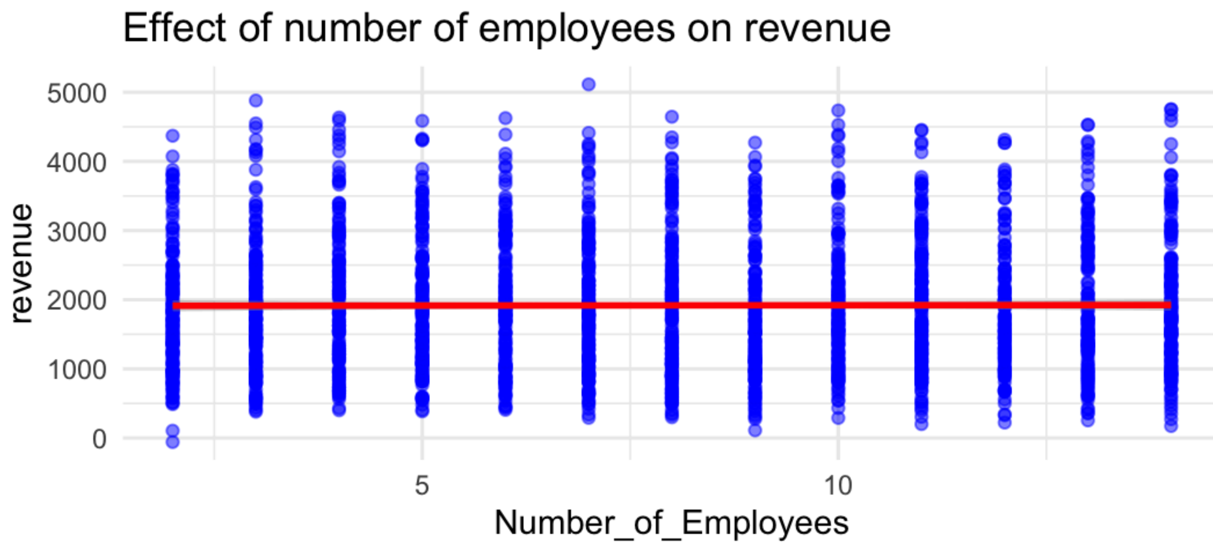
Data Insights:

What makes this data set interesting is the marketing spend attribute. Even though the coffee shop potentially spends more on marketing for that day, it doesn't directly correlate into increased revenue because the type of marketing is important as well. Based on numbers alone, we are unsure what type of strategies were used or if the intended marketing tactic reached its audience. A higher revenue day generally has higher marketing spend but it is not a guarantee. If the advertisement does not reach its audience then there won't be a difference in number of customers and average order value. A greater increase in marketing spending has a higher chance to reach its audience which is why a higher marketing spending generally leads to higher revenue due to increase in customers. However, the inverse is also true where a lower marketing spending day could still have a high number of customers.




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


Effect of operating hours per day on revenue
revenue
5000
4000
3000
2000
1000
0
7.5
10.0
12.5
15.0
17.5
Operating_Hours_Per_Day



Effect of number of employees on revenue
revenue
5000
4000
3000
2000
1000
0
5
10
Number_of_Employees

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

Opportunities to Use This Data Set:

This dataset could be used to predict revenue based on certain internal and external factors for a coffee shop. The coffee shop can also use this as an analysis to see where they can cut costs. Since this dataset shows that operating hours, number of employees, and foot traffic do not have much impact, they can find ways to reduce costs from these areas. A coffee shop can also look to find ways to increase their average order value to increase profits.

<u>Data Set 2: [Advertisement - Click on Ad dataset]</u>

Data Structure:

The data consists of one table. Use Ad Topic Line and Timestamp as a composite primary key to ensure data uniqueness.

Field Name	Data Type	Description (Optional)
Daily Time Spent on Site	Number	Time spent by the user on the website per day (in minutes).
Age	Number	User's age.
Area Income	Number	Average income of the user's residing area.
Daily Internet Usage	Number	Time spent online per day (in minutes).
Ad Topic Line	Short Text	The theme of the advertisement.
City	Short Text	The city where the user is located.
Male	Number	User's gender (1 for male, 0 for female).
Country	Short Text	The country where the user is located.
Timestamp	Date/Time	The timestamp of the ad display.
Clicked on Ad	Number	Whether the user clicked on the ad (1 for clicked, 0 for not clicked).

Analysis with SQL and R:

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

```
Call:
glm(formula = Clicked.on.Ad ~ Daily.Time.Spent.on.Site + Age +
    Area.Income, family = binomial, data = df)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.504e+01  1.443e+00  10.420   <2e-16 ***
Daily.Time.Spent.on.Site -2.048e-01  1.565e-02 -13.085   <2e-16 ***
Age                       1.630e-01  1.785e-02   9.132   <2e-16 ***
Area.Income              -1.173e-04  1.274e-05  -9.206   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.29  on 999  degrees of freedom
Residual deviance:  391.98  on 996  degrees of freedom
AIC: 399.98
```

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

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

Data Trends:

When analyzing the data’s most common and grouped information, we can see that lower-income groups have a higher click rate on ads. The highest engagement (100%) is seen in the Low-Income group, while the Most High-Income group has the lowest engagement (~27.6%). As income increases, ad engagement tends to decrease.

Even though income is said to have a significant effect, it does not have a direct effect on ad clicks in a linear fashion. While lower-income users show the highest engagement, there is variability in engagement among middle and high-income groups. This suggests that other factors, such as ad relevance and user preferences, may also play a role.

For the effect of income on ad click probability, a downward trend is observed. As income increases, the probability of clicking on an ad decreases. This indicates that users from wealthier backgrounds may be less influenced by online ads, potentially due to different purchasing behaviors or ad fatigue.

Similarly, when analyzing daily time spent online and ad clicks, users who spend less time online (Short Time Spent group) have the highest ad click rate (~99.14%). Those with moderate time spent online (~70% ad click rate) engage significantly but less than the first group. Meanwhile, long-time online users have the lowest click rate (~11.9%), suggesting they may be less interested in ads.

The effect of online time on ad click probability shows a clear inverse relationship. Users who spend more than 70 minutes online have a drastically lower probability of clicking on ads. Based on these data trends, it seems that a combination of lower income and shorter online time results in the highest ad engagement.

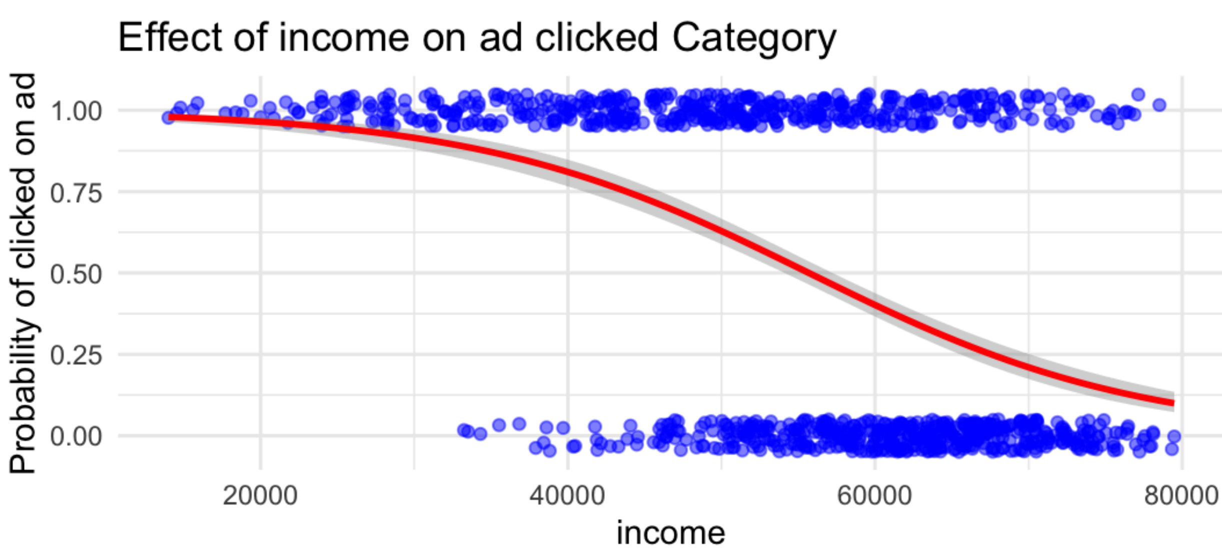
Data Insights:

The reason for choosing this dataset is that it shows how people from different income groups and who spend different amounts of time online interact with online ads, thereby giving advertisers some inspiration to create different content for different groups of people.

SQL 1* SQL 2* SQL 3*

```sql
SELECT
    Age,
    CASE
        WHEN "Area Income" < 20000 THEN 'Low Income'
        WHEN "Area Income" BETWEEN 20001 AND 40000 THEN 'Middle Income'
        WHEN "Area Income" between 40001 and 60000 then 'High Income'
        ELSE 'Most High Income'
    END AS Income_Group,
    AVG("Clicked on Ad") * 100 AS Click_Rate
FROM advertising
```

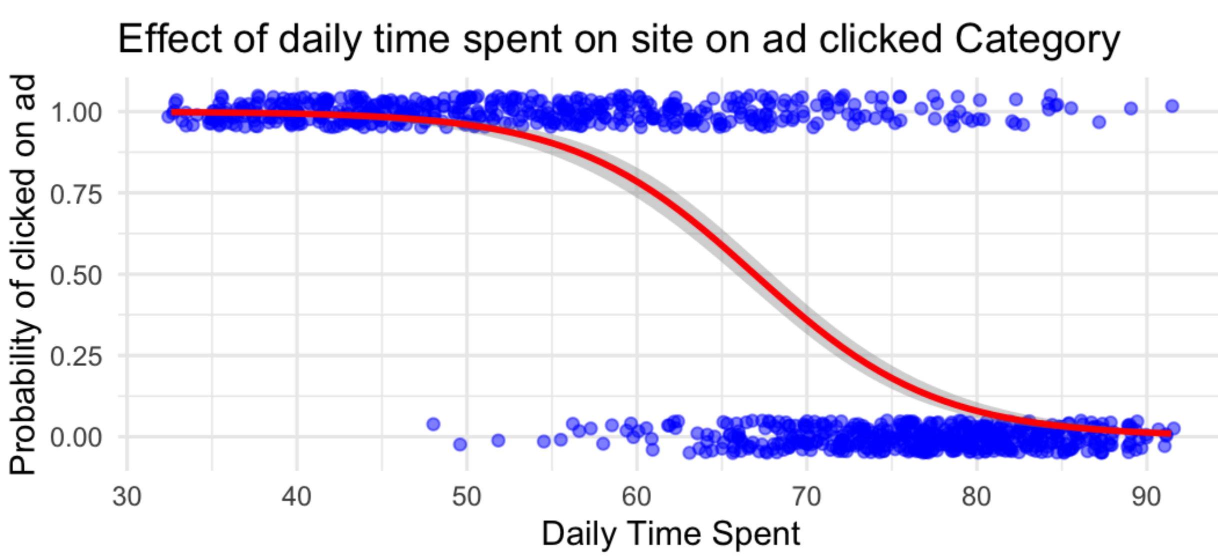
	Age	Income_Group	Click_Rate
1	26	High Income	57.1428571428571
2	39	Low Income	100.0
3	48	Middle Income	92.6470588235294
4	35	Most High Income	27.6190476190476



SQL 1* SQL 2* SQL 3* SQL 5*

```sql
SELECT
    CASE
        WHEN "Daily Time Spent on Site" < 50 THEN 'Short Time Spent'
        WHEN "Daily Time Spent on Site" BETWEEN 50.01 AND 70 THEN 'Moderate Time Spent'
        ELSE 'Long Time Spent'
    END AS Time_Spent_Group,
    COUNT(*) AS User_Count,
    SUM("Clicked on Ad") AS Ad_Clicks,
    ROUND(SUM("Clicked on Ad") * 100.0 / COUNT(*), 2) AS Ad_Click_Rate
FROM advertising
GROUP BY Time_Spent_Group
ORDER BY Time_Spent_Group;
```

	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



Opportunities to Use This Data Set:

Marketing teams can analyze whether higher-income areas have higher ad click rates and adjust targeting accordingly.

By correlating online time with ad clicks, businesses can optimize the best time to run ads, not only that, businesses can also modify ad pricing and delivery.

Optimize audience targeting by understanding which regions or income groups click on ads more frequently.

Identifies whether longer browsing times lead to more ad clicks and potential purchases.

<u>Data Set 3: [Food App Business]</u>

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.

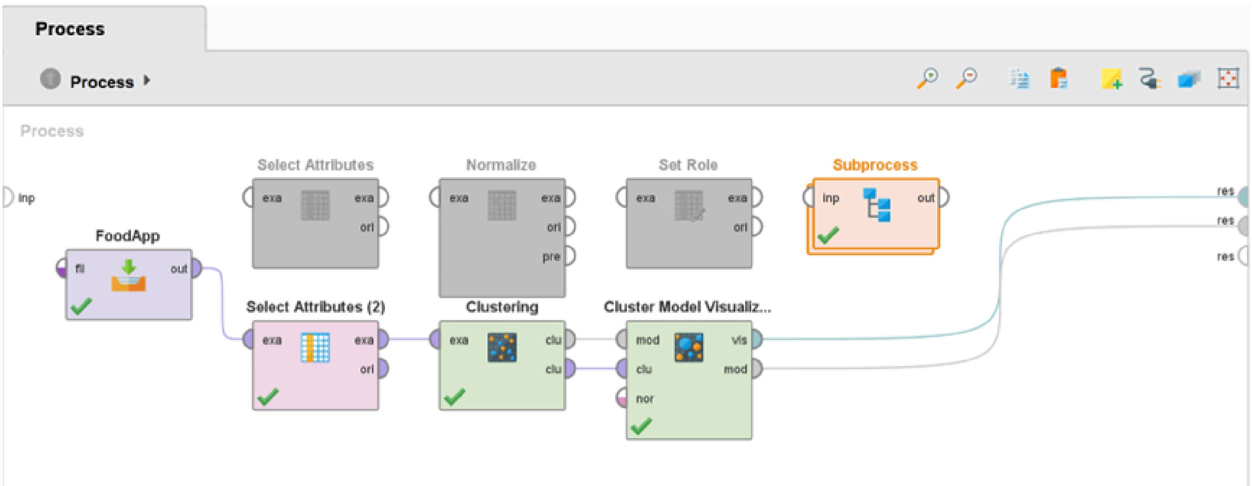
FoodAppBusiness-PKadded		CREATE TABLE "FoodAppBusiness-PKadded" ("CustomerId" INTEGER, "MonthlyIncome" INTEGER, "ActiveSinceDays" INTEGER, "Age" INTEGE
CustomerId	INTEGER	"CustomerId" INTEGER
MonthlyIncome	INTEGER	"MonthlyIncome" INTEGER
ActiveSinceDays	INTEGER	"ActiveSinceDays" INTEGER
Age	INTEGER	"Age" INTEGER
Graduate	INTEGER	"Graduate" INTEGER
Married	INTEGER	"Married" INTEGER
Single	INTEGER	"Single" INTEGER
NoOfChildren	INTEGER	"NoOfChildren" INTEGER
NoOfTeenager	INTEGER	"NoOfTeenager" INTEGER
NoOfDaysSinceLastPurchase	INTEGER	"NoOfDaysSinceLastPurchase" INTEGER
AmountSpendOnWines	INTEGER	"AmountSpendOnWines" INTEGER
AmountSpentOnFruits	INTEGER	"AmountSpentOnFruits" INTEGER
AmountSpentOnMeat	INTEGER	"AmountSpentOnMeat" INTEGER
AmountSpentOnFish	INTEGER	"AmountSpentOnFish" INTEGER
AmountSpentOnSweet	INTEGER	"AmountSpentOnSweet" INTEGER
AmountSpentOnGold	INTEGER	"AmountSpentOnGold" INTEGER
NoOfDealsWithDiscount	INTEGER	"NoOfDealsWithDiscount" INTEGER
NoOfWebPurchase	INTEGER	"NoOfWebPurchase" INTEGER
NoOfCatalogPurchase	INTEGER	"NoOfCatalogPurchase" INTEGER
NoOfStorePurchase	INTEGER	"NoOfStorePurchase" INTEGER
NoOfWebVisitsMonth	INTEGER	"NoOfWebVisitsMonth" INTEGER
PurchasedIn1stCampaign	INTEGER	"PurchasedIn1stCampaign" INTEGER
PurchasedIn2ndCampaign	INTEGER	"PurchasedIn2ndCampaign" INTEGER
PurchasedIn3rdCampaign	INTEGER	"PurchasedIn3rdCampaign" INTEGER
PurchasedIn4thCampaign	INTEGER	"PurchasedIn4thCampaign" INTEGER
PurchasedIn5thCampaign	INTEGER	"PurchasedIn5thCampaign" INTEGER
TotalNoOfCampaignAccepted	INTEGER	"TotalNoOfCampaignAccepted" INTEGER
CustomerComplain	INTEGER	"CustomerComplain" INTEGER

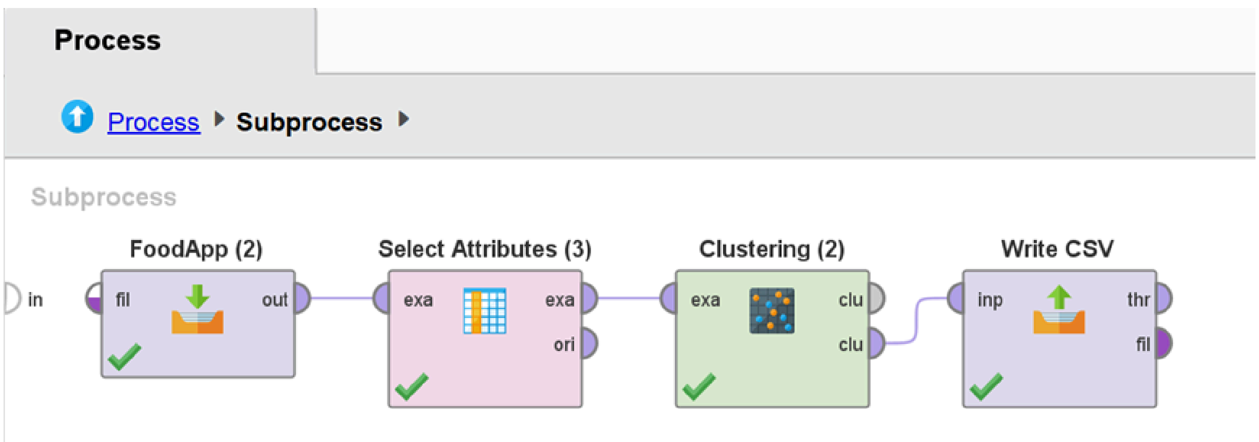
Analysis with RapidMiner and SQL:

RapidMiner was used to perform K-Means clustering on the food app data to identify customer segments. The 2205 observations were divided into 6 clusters to analyze patterns and insights. The clustering analysis revealed that income, number of children, and spending on categories such as wine and meat were the key differentiators between clusters.

The operator Cluster Model Visualization was used to generate a heatmap to help with understand the outcome of clustering analysis along with the centroid table.

A new CSV file with cluster group number was created in the subprocess so that further analysis could be carried out with SQL. Trends within each cluster, such as how promotional campaigns affect spending, were investigated.





Data Trends:

The centroid table shows the average characteristics of each cluster across different features. In general, Cluster 0 has the highest average income ($83,821.79) and a preference for premium products. They make more web purchases and have low reliance on discounts, showing high engagement with marketing campaigns.

Cluster 1 consists of price-sensitive individuals with low income who tend to make more in-store visits. They are less responsive to promotional offers.

Cluster 2 has moderate income and exhibits balanced spending across product categories. They make web purchases less frequently, spend more on meat, and have a moderate response to marketing campaigns.

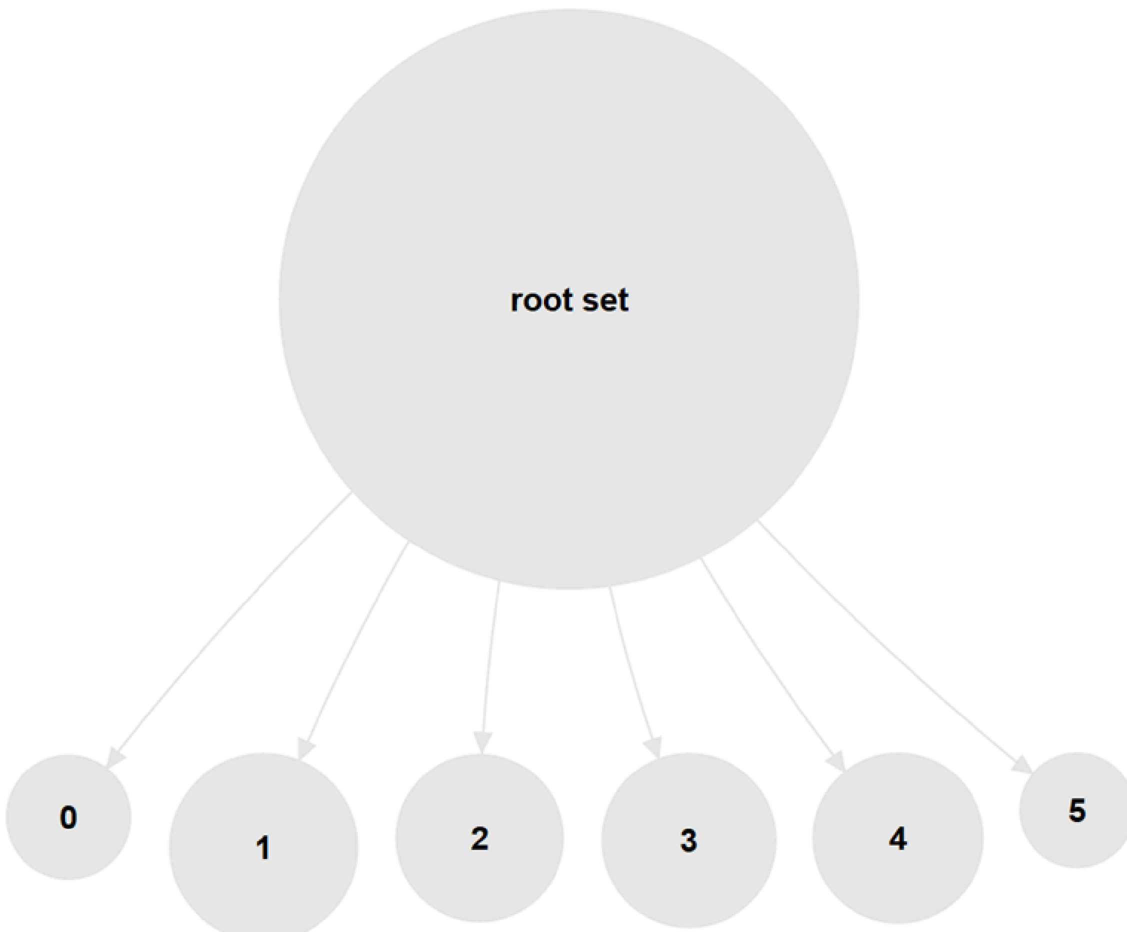
Cluster 3 has moderate income and prefers in-store visits over online transactions. They are more discount-focused, with a lower frequency of web purchases and minimal response to online promotions.

Cluster 4 is similar to Cluster 0, but they also spend moderately in stores.

Cluster 5 has the lowest average income ($18,939.30). They heavily engaged with discounts when ordering. They have limited spending across all product categories and minimal online purchases.

Based on the characteristics of the clusters, higher-income clusters tend to make more web purchases. Additionally, the number of marketing campaigns accepted and spending on discounts had a noticeable effect on the purchasing behavior of certain clusters.

According to the result set generated from the SQL query, promotions can always stimulate orders but are much less effective when applied to customers with budget constraints. The relationship between promotions and order numbers is non-linear.



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

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

FoodAppBusiness-clustered		CREATE TABLE "FoodAppBusiness-clustered" ("Monthl
MonthlyIncome	INTEGER	"MonthlyIncome" INTEGER
Age	INTEGER	"Age" INTEGER
NoOfChildren	INTEGER	"NoOfChildren" INTEGER
NoOfTeenager	INTEGER	"NoOfTeenager" INTEGER
AmountSpendOnWines	INTEGER	"AmountSpendOnWines" INTEGER
AmountSpentOnFruits	INTEGER	"AmountSpentOnFruits" INTEGER
AmountSpentOnMeat	INTEGER	"AmountSpentOnMeat" INTEGER
AmountSpentOnFish	INTEGER	"AmountSpentOnFish" INTEGER
AmountSpentOnSweet	INTEGER	"AmountSpentOnSweet" INTEGER
NoOfDealsWithDiscount	INTEGER	"NoOfDealsWithDiscount" INTEGER
NoOfWebPurchase	INTEGER	"NoOfWebPurchase" INTEGER
NoOfStorePurchase	INTEGER	"NoOfStorePurchase" INTEGER
NoOfWebVisitsMonth	INTEGER	"NoOfWebVisitsMonth" INTEGER
TotalNoOfCampaignAccepted	INTEGER	"TotalNoOfCampaignAccepted" INTEGER
id	INTEGER	"id" INTEGER
cluster	TEXT	"cluster" TEXT

```
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

Data Insights:

If the advertisement does not reach its audience then there won't be a difference in number of customers and average order value. A greater increase in marketing spending has a higher chance to reach its audience which is why a higher marketing spending generally leads to higher revenue due to increase in customers. However, the inverse is also true where a lower marketing spending day could still have a high number of customers.

The most unexpected outcome identified by the analysis was that the number of campaigns accepted did not have a strong relation with spending, regardless of the content of each campaign. Spending habit can be affected by promotions, but not always. Monthly income is in a higher hierarchy of purchases. By analyzing key features such as the number of web and store purchases and some other attributes, we can divide customers into groups. Businesses can pay more attention to the high-engaged groups, and based on their characteristics, identify trends of each group to increase customer retention and purchasing decisions optimally.

We observe that customers who react the most frequently to promotions or discounts are more likely to place orders. This indicates the importance of targeted marketing strategy in driving sales. The dataset also highlights the varying preferences in the amount spent on each category of food across different customer segments, which can also help businesses tailor menus and marketing strategies for specific demographic groups, for example, price-sensitive customers, who value the price first, to optimize revenue.

Opportunities to Use This Data Set:

We have only inspected a subset of the attributes. There are other features that can be extracted from the dataset, depending on the target audience and business strategy of this industry, based on the life philosophy nowadays.

This dataset can also be leveraged to drive models to assess business decisions, such as improving inventory management and balancing the labor of in-store and web.

From our investigation, businesses can get a better understanding on their current, existing customers and tailor their marketing strategies for each cluster, along with an optimization algorithm.

To increase revenue and customer retention, businesses can target the highest-spending segment (e.g. Cluster 0) with personalized promotions. Designing loyalty programs for repeat orders can attract customers with a relatively large number of children and teenagers (Cluster 1 and Cluster 5) since these customers may have a spending priority on necessities.

Additionally, when a significant budget for fruit and fish is allocated, these customers focus more on high-quality or healthy food. Then, developing new healthy choices can attract them. Also, labeling the calorie content, and highlighting the use of organic ingredients in advertisements can increase their order numbers and size.

For customers who tend to make in-store orders, dispensing in-store promotions and targeted discounts can grow conversion rates.

Analyzing the attribute, ActiveSinceDays, in combination of some external factors, can be used to assess if businesses should improve the app functionality along with promotions, for instance, daily sign-in rewards, and ease of navigation (e.g. personalized recommendations) to keep customers engaged with the app constantly, can increase the likelihood of them placing orders.

Section 3: Conclusion

Data Set 1: [Coffee Shop Revenue]

This dataset highlighted important insights into sales performance based on internal and external factors. The rapid miner linear regression revealed a positive correlation between the number of customers, average order size, and marketing spending, and daily revenue. Hours of operation, number of employees, and location foot traffic did not play a significant role in daily revenue. Though marketing spending was a significant variable and generally led to more revenue, it was not a direct result of it. Marketing spending also did not directly affect customers, average order size, and location foot traffic. This would help coffee shops make data-driven decisions on where to make improvements and cut costs to generate the most revenue.

Data Set 2: [Advertisement - Click on Ad dataset]

This dataset highlighted crucial insights into online advertisement engagement based on user demographics and online behavior. The logistic regression analysis revealed significant relationships between income levels, daily internet usage, and ad-click probability. Lower-income users showed the highest engagement, while higher-income users were less likely to click on ads. Additionally, users who spent less time online had the highest click rates, whereas long-time internet users had significantly lower engagement. These findings help advertisers and marketers optimize ad placements, refine audience targeting, and allocate marketing budgets effectively to maximize engagement.

Data Set 3: [Food App Business]

The dataset emphasized the importance of understanding customers' spending patterns and tailoring marketing strategies accordingly to maximize sales and customer retention. The clustering groups customers into distinct segments based on similar behaviors, preferences, or demographics, by comparing values in the dataset with the means of k number of groups.

The moderately engaged segment is less price-sensitive. They are less likely to be influenced by promotions. While high-income individuals prefer online purchases, they are highly responsive to promotional campaigns that focus on cost-effectiveness, taking advantage of high-quality offers or exclusive deals. Individuals with relatively low income, prefer in-store orders and care more about the amount of savings and absolute price, rather than the value of products.

By analysis on the clusters, businesses can design more targeted and effective marketing campaigns or business strategies that resonate with the most customer groups, leading to higher activity, increased sales or spending, and improved customer loyalty.

References

Data Set 1: Linear Regression
https://www.kaggle.com/datasets/himelsarder/coffee-shop-daily-revenue-prediction-dataset
Data Set 2: Decision Tree
https://www.kaggle.com/datasets/gabrielsantello/advertisement-click-on-ad
Data Set 3: Clustering
https://www.kaggle.com/datasets/ybifoundation/food-app-business