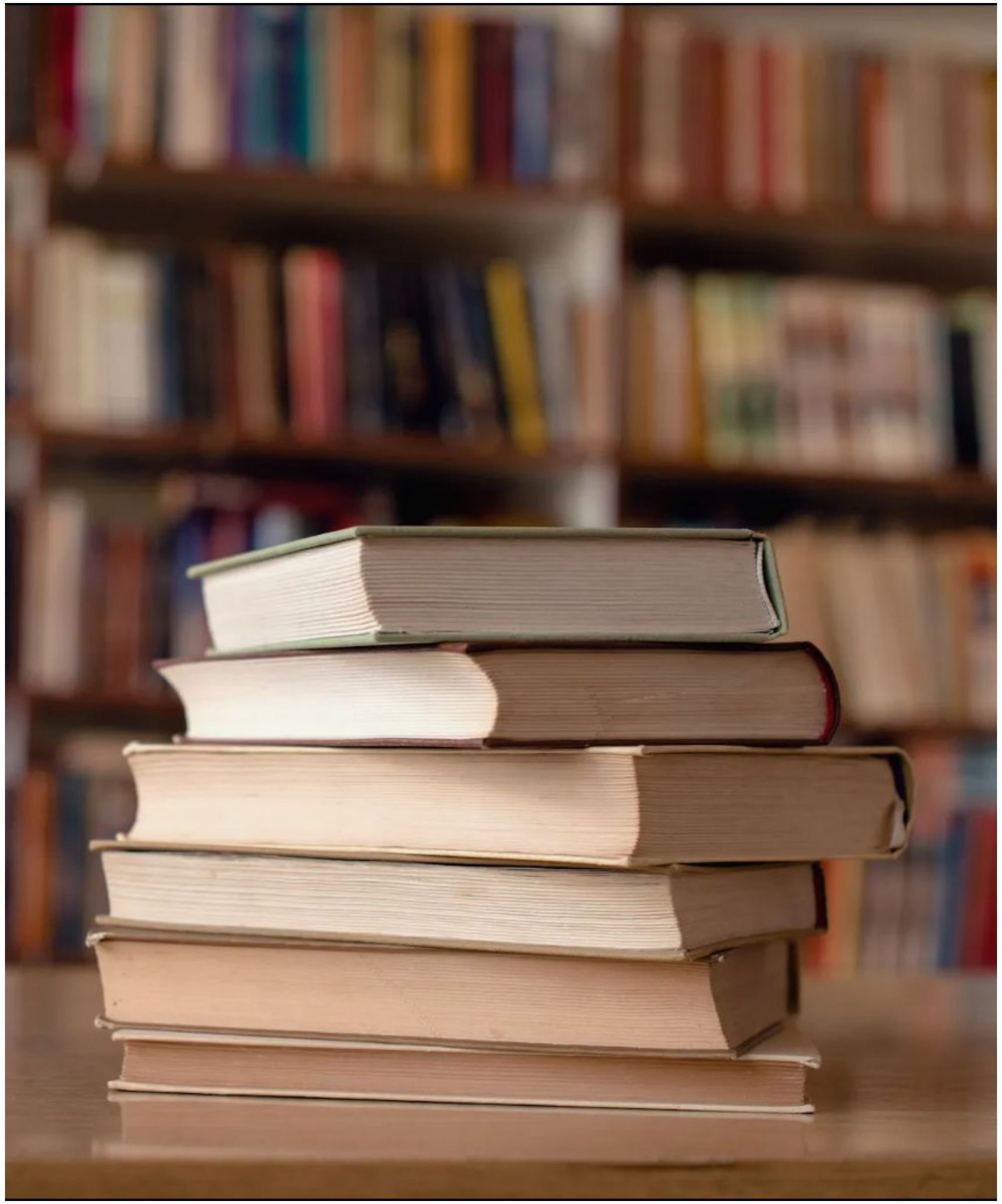
Powell's Bookstore Client DATA ANALYSIS (2024)



Content

- **Business Case** –

  1. Monthly Spending Prediction
  The goal is to forecast how much a customer is likely to spend each month. This leads to understanding purchasing patterns and create personalized/localized offers.

  2. eBook Subscription Likelihood
  The second part focuses on predicting the likelihood of a customer subscribing to an eBook service. This helps businesses identify potential subscribers and target them with tailored marketing campaigns.

- **Data Acquisition** – Customer data was obtained in CSV format.
- **Data Preparation** – Unnecessary data was dropped, multiple data collected as one column was separated, and the age was calculated for better representation of data.
- **Data Visualization** – Graphs were created using Tableau and Sweetviz to understand the relation of features.

Data Preparation

- **Total number of records** = 16,519
- **Following missing values were found**:
  - Title (88 records), Middle Name(9534 records), Suffix(2 records), Street address2(276 records)
- **Missing values treatment:**
  - The values with more than 20% missing and with no significance to the analysis were dropped.
- **Birth date value modification:**
  - Birth date(MM/DD/YYYY) : The value was calculated and modified to show the actual age of the clients.
- **City-ZipCode-State value modification:**
  - The column was divided into three separate columns to better represent significant relationship.
- **eBook Subscriber Flag remapping:**
  - The data was remapped to show ‘No’ for 0 and ‘Yes’ for 1.
- **Homeowner Status Flag remapping:**
  - The data was remapped to show ‘No’ for 0 and ‘Yes’ for 1.

Data Preparation

Summary table

	Variable class	# unique values	Missing observations	Any problems?
Education Level	character	5	0.00 %	
Occupation	character	5	0.00 %	
Gender	character	2	0.00 %	
Marital Status	character	2	0.00 %	
Home Owner Status	character	2	0.00 %	
Number of Cars Owned	numeric	5	0.00 %	
Number of Children At Home	numeric	6	0.00 %	
Total Number of Children	numeric	6	0.00 %	
Annual Income	numeric	15482	0.00 %	
Avg Monthly Spend	numeric	152	0.00 %	×
eBook Subscriber Flag	character	2	0.00 %	
Age	numeric	70	0.00 %	×
City	character	77	0.00 %	
ZipCode	numeric	77	0.00 %	
State	character	33	0.00 %	

- The Summary Table for features were collected by ***DataReporter*** and ‘Avg Monthly Spend’ and ‘Age’ had red flags.
- ‘Avg. Monthly Spend’: While the median value was 68, the following possible outlier values were detected: "146", "147", "148",. . . , "172", "175", "176" (21 values).
- Age: While the median value was 61, possible outliers under 10 were detected and were omitted for data analysis.

Exploratory Data Analysis

The Features (i.e., variables) are segregated into three Categories namely:

- **Dependent Variable (Target) –** Two variables were to be predicted: Monthly Spend per client and the likelihood of eBook Subscription
- **Noise Features :** Variables which would not have a significant impact on the value of the Target

    Redundant

    Feeder variables

- **Predictor Variables** : Variables which were considered to have an impact on the Target

Exploratory Data Analysis (contd..)

Noise Variables

Title
First Name
Middle Name
Suffix
Street Address 1
Street Address 2
Customer ID

Predictor Variables

Age	Total number of children
Education Level	Annual Income
Occupation	Average Monthly Spend
Gender	City
Is Magnet	Zip Code
Marital Status	State
Homeowner Status	
Number of Cars Owned	
Number of Children at home	

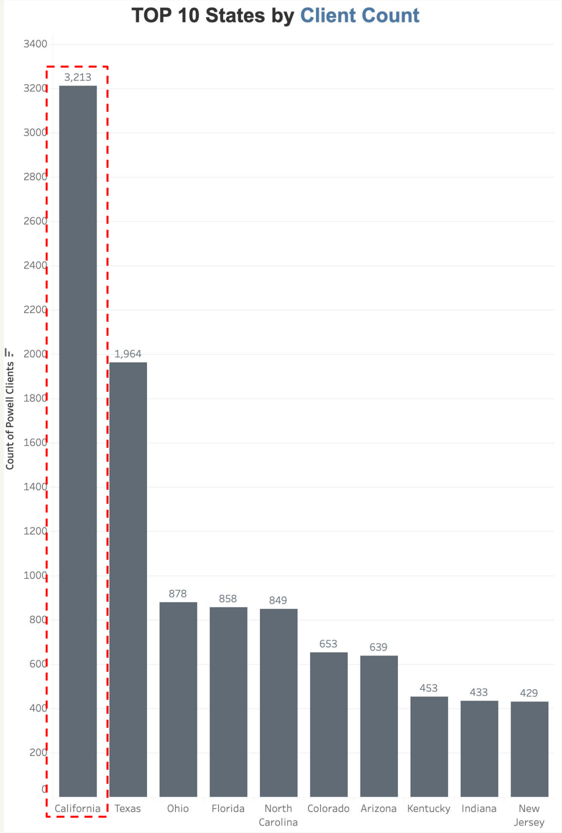
Data Visualization

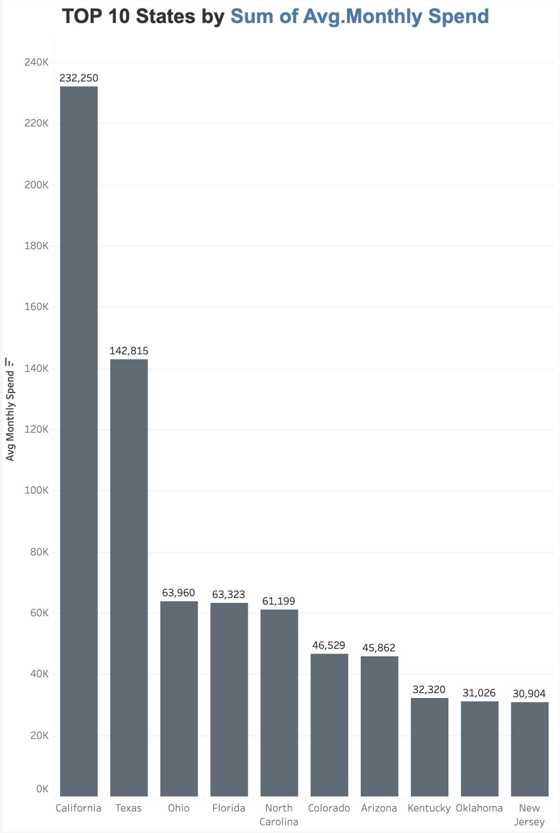
- Data visuals were created using ***Tableau*** where charts are plotted using the independent variables against the dependent variable (Monthly Spend/eBook subscription flag).

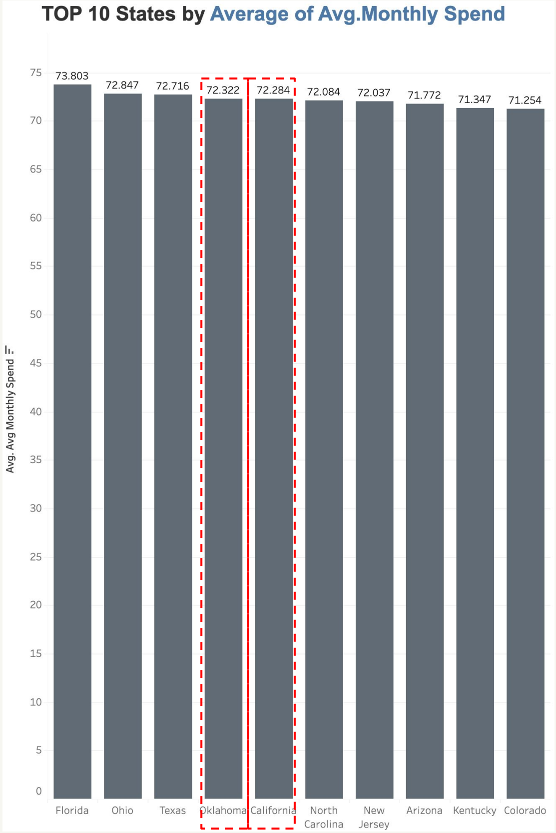
- Reports were created using ***DataReporter*** and ***DataExplorer*** to understand the correlation and aggregation of features.

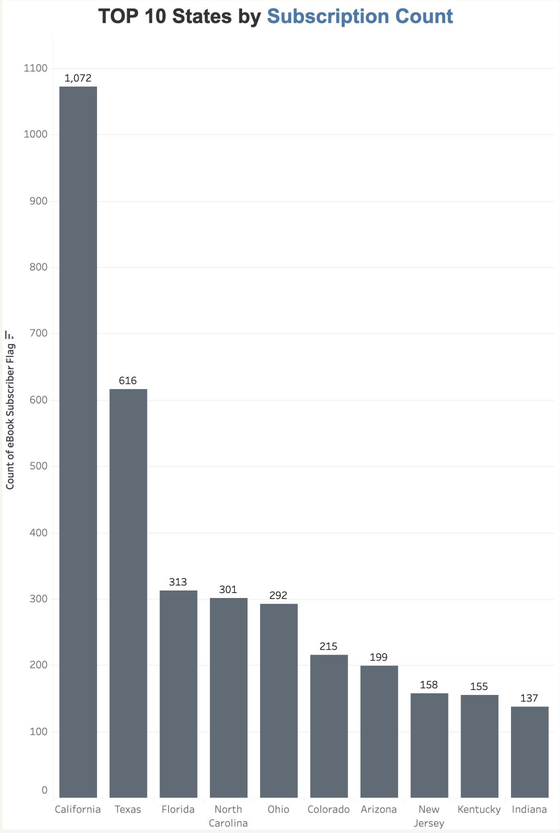
- These charts help us in getting a preliminary idea about the **relationship** between the **independent variables** and the **dependent variable.**

Data Visualization





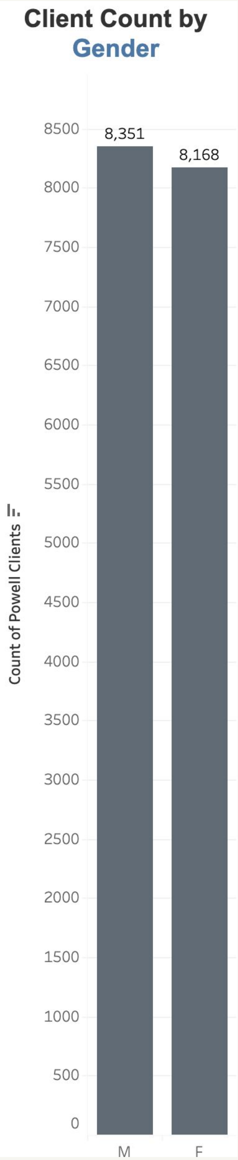


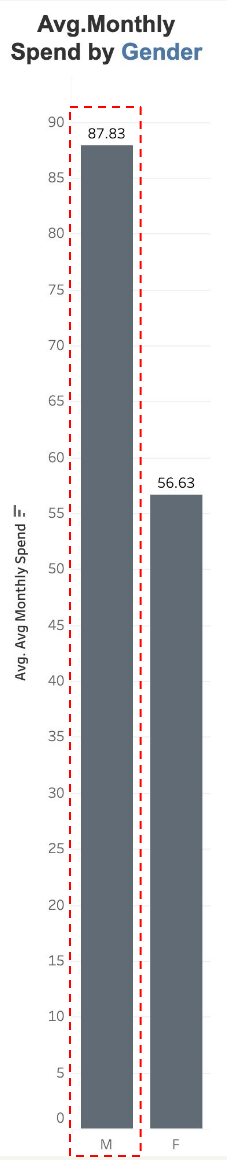


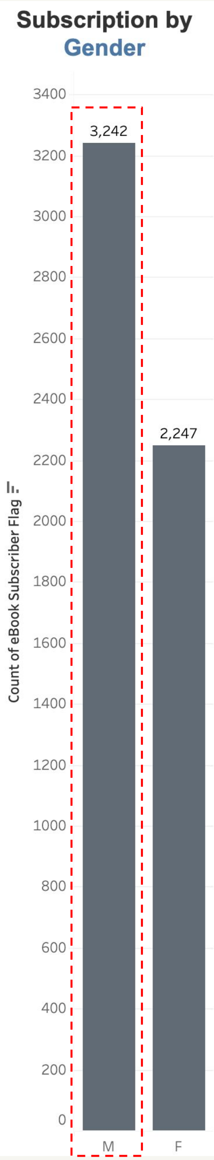
Key Highlights

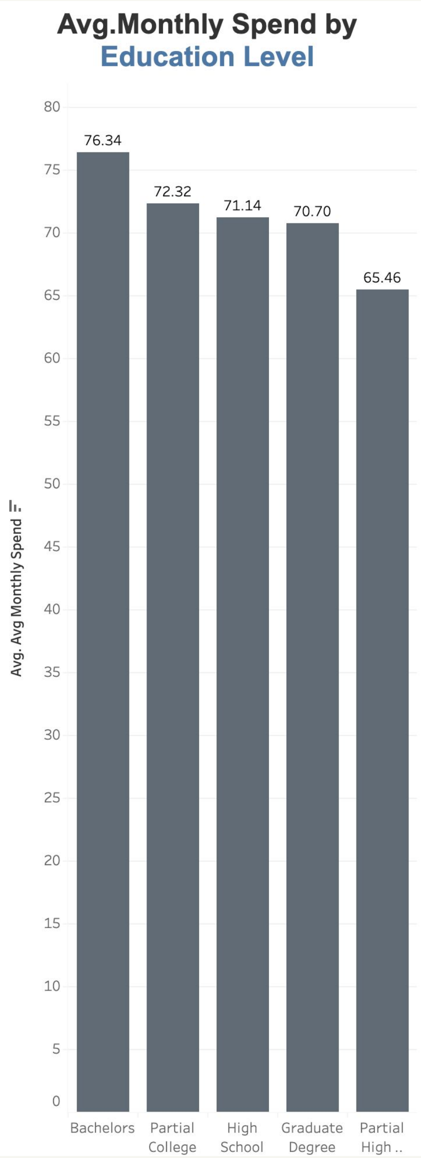
- ➢ **California** ranks **#1** in total clients, subscribers, and monthly spending, with figures nearly twice those of the #2 ranked states in all three categories. However, it ranks only **5th** in average monthly spend per client.
- ➢ **Oklahoma,** despite not making the TOP 10 in client count or subscription count, ranks **4th** in average monthly spend.
- ➢ While there are notable gaps between the top two states and the rest in terms of client count, total monthly spend, and subscription count, **the average of ‘Average monthly spend’ across the top 10 states shows no significant difference.**

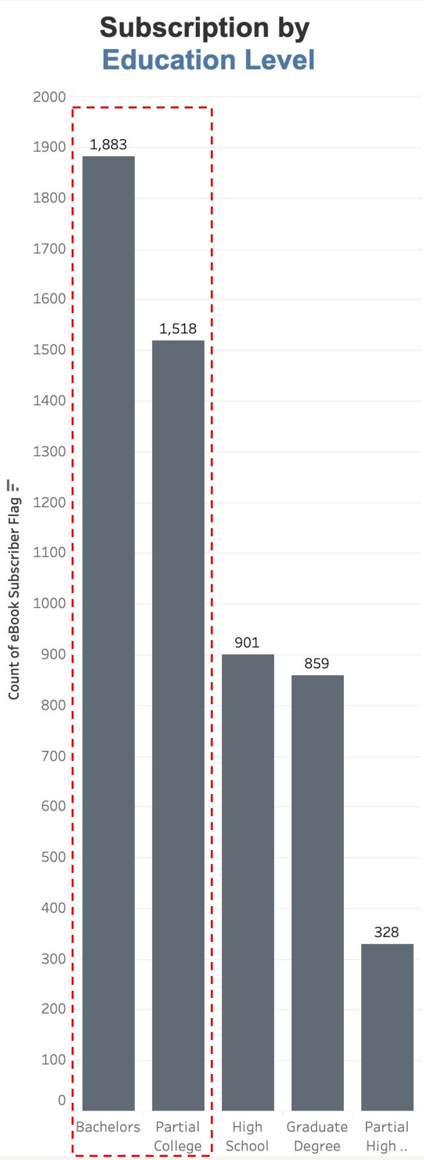
Data Visualization









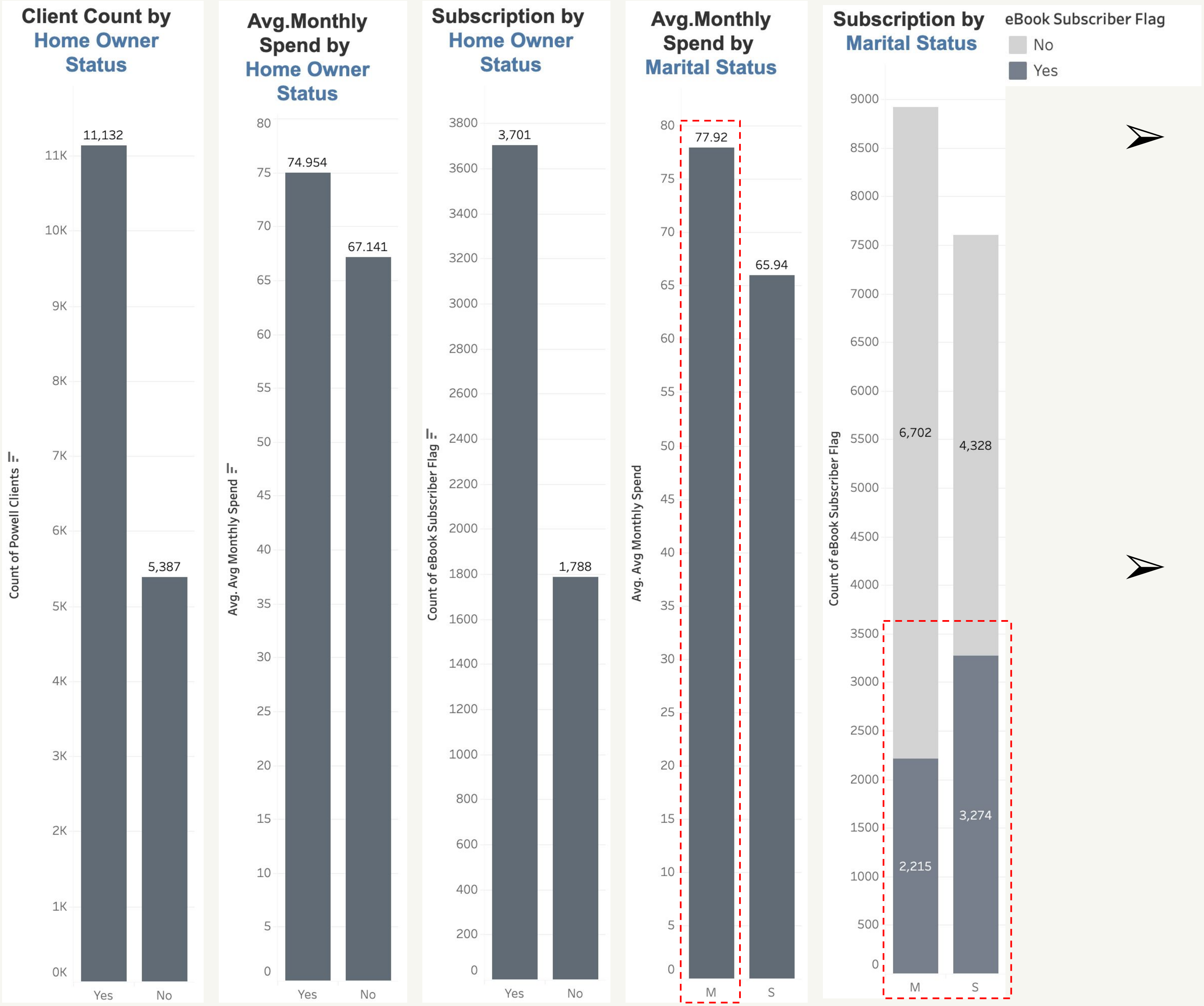


Key Highlights

- ➢ There is no significant gender difference in total client count, but **males** show a **higher average monthly spend** (55% more) and **higher subscription status** (44% more) than females.

- ➢ **Education level** rankings for both average monthly spend and subscription status are identical. However, **the gap in subscription status is more significant**, with clients holding bachelor's degrees or partial college education accounting for **60%** of the subscription group.

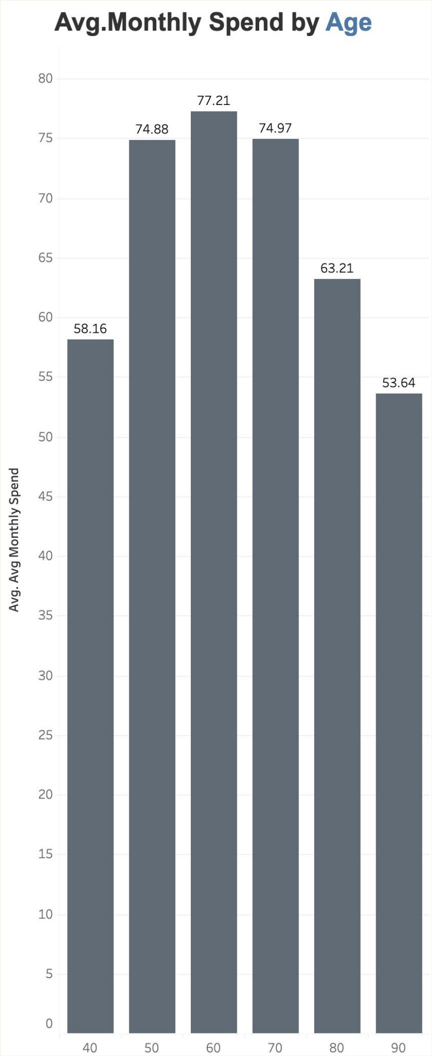
Data Visualization

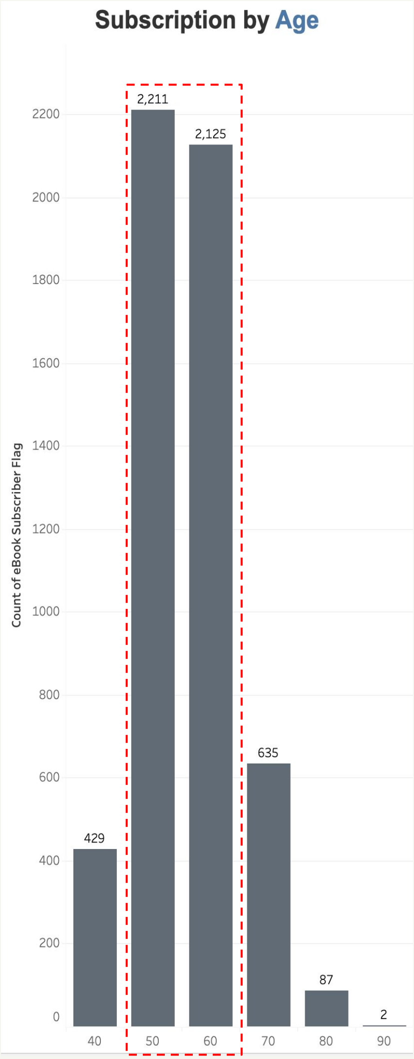


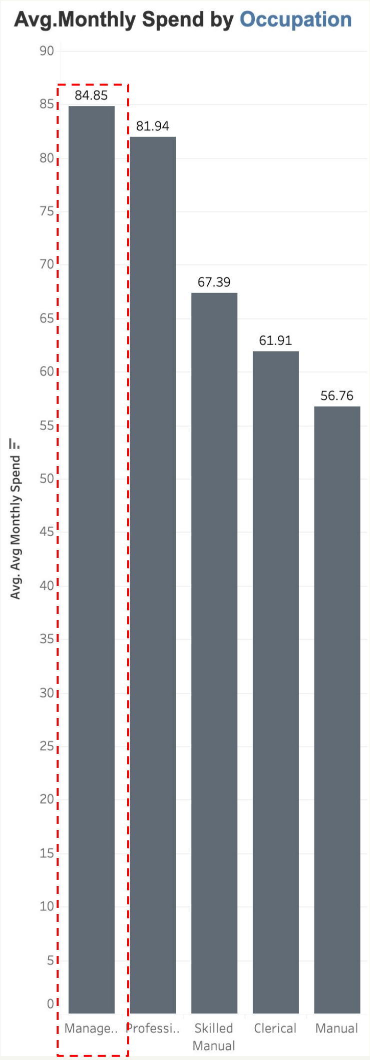
Key Highlights

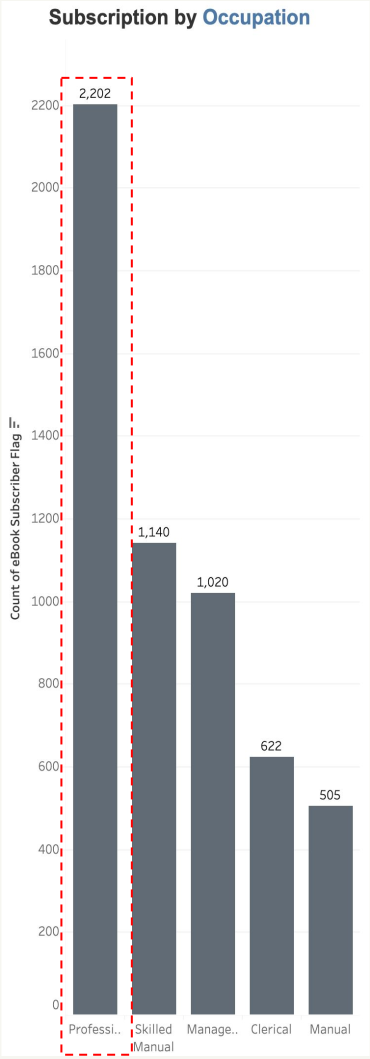
- ➢ For home ownership status, **67%** of total clients are **homeowners**. Homeowners lead both in average monthly spend and subscription, with a particularly strong lead in subscriptions—**107%** higher than non-homeowners.
- ➢ Regarding marital status, **married** clients had **higher average spending** and a **larger client count**. *Interestingly however,* ***singles*** *led in* ***subscription count***.

Data Visualization









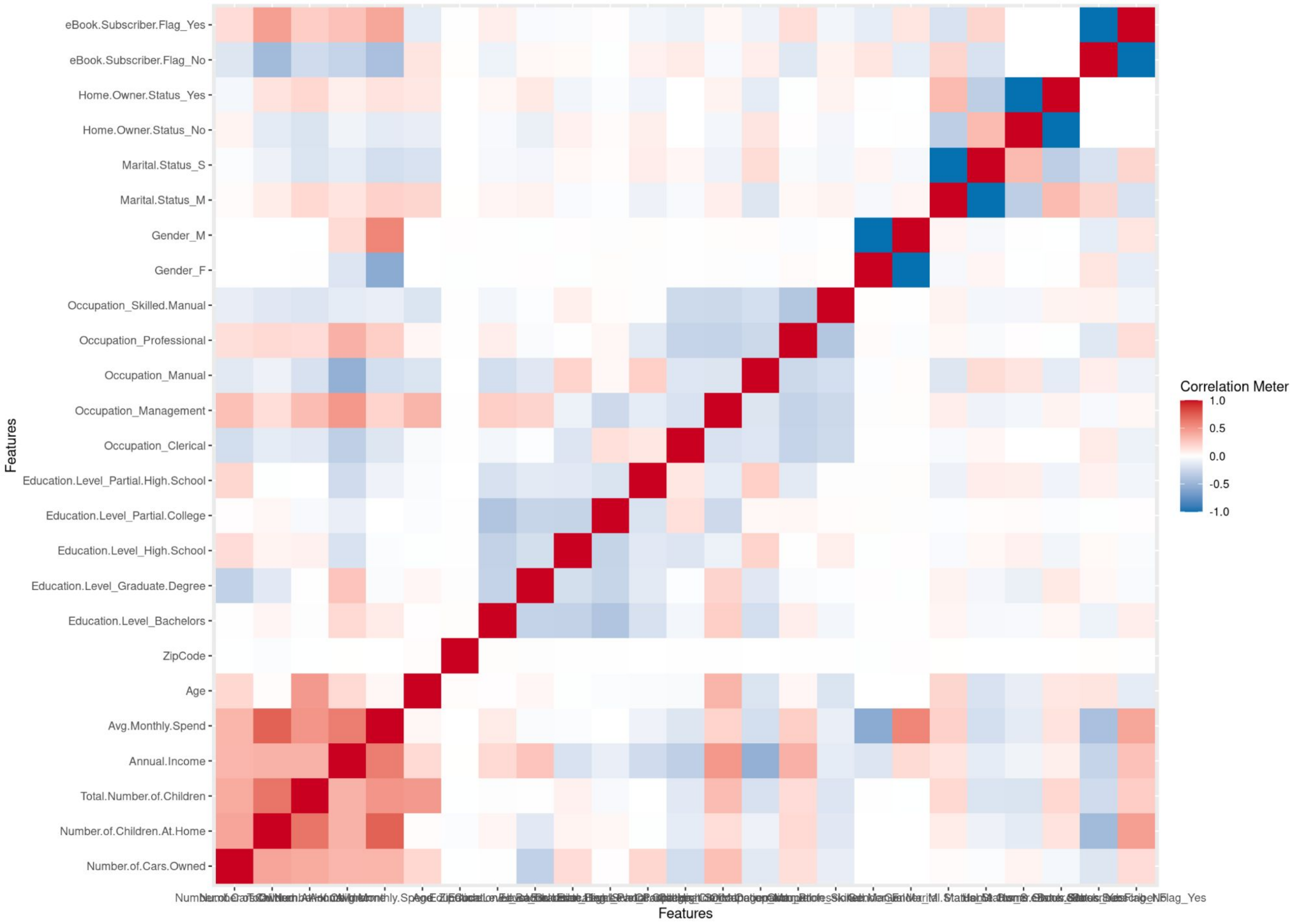
Key Highlights

- In terms of **subscription status** by age, the **50's and 60's** age group is dominant, accounting for over **70%** of all subscribers. In contrast, the **40's and 70's** age groups have **relatively low subscriber counts** despite having higher average monthly spend.
- **'Managers'** lead in **average monthly spend**, but rank only **#3** in **subscriber count**. The '**Professional**' occupation is the clear leader in subscribers, with **93%** more subscribers than 'Skilled Manual' occupation.

Data Visualization

Correlation Analysis

```
## 2 features with more than 20 categories ignored!
## City: 77 categories
## State: 33 categories
```



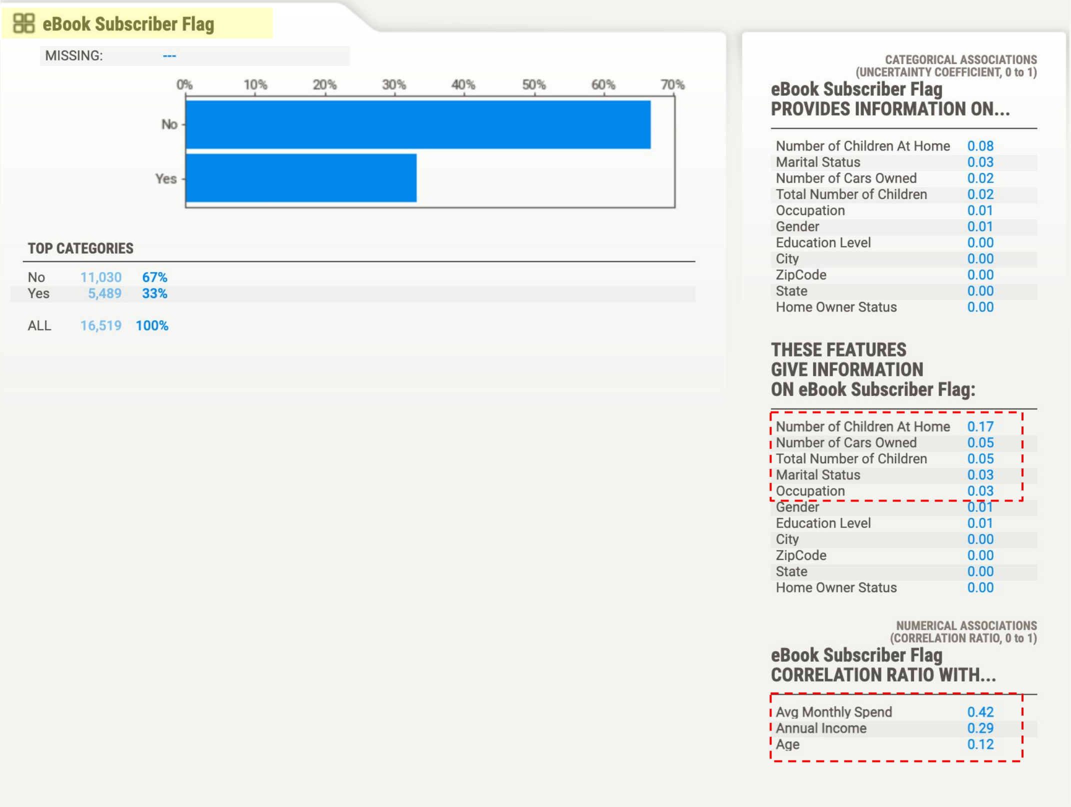
Key Highlights

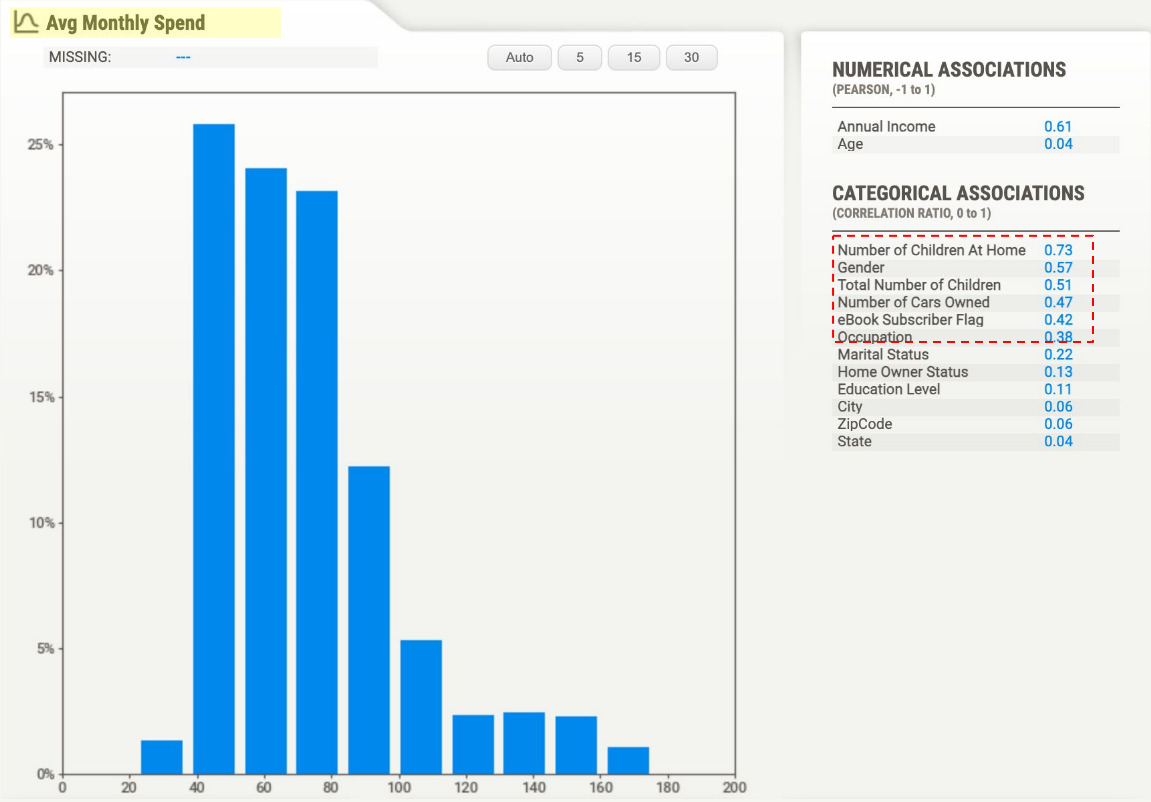
The ***DataExplorer*** report indicated the following correlation analysis:

➢ '**Avg monthly spend**' had a strong correlation with the '**number of children at home**', '**annual income**', '**gender**', 'total number of children' and 'number of cars owned'.

➢ '**Subscription**' had a strong relation with '**number of children at home**','total number of children', 'annual income' and 'number of cars owned'.

Data Visualization

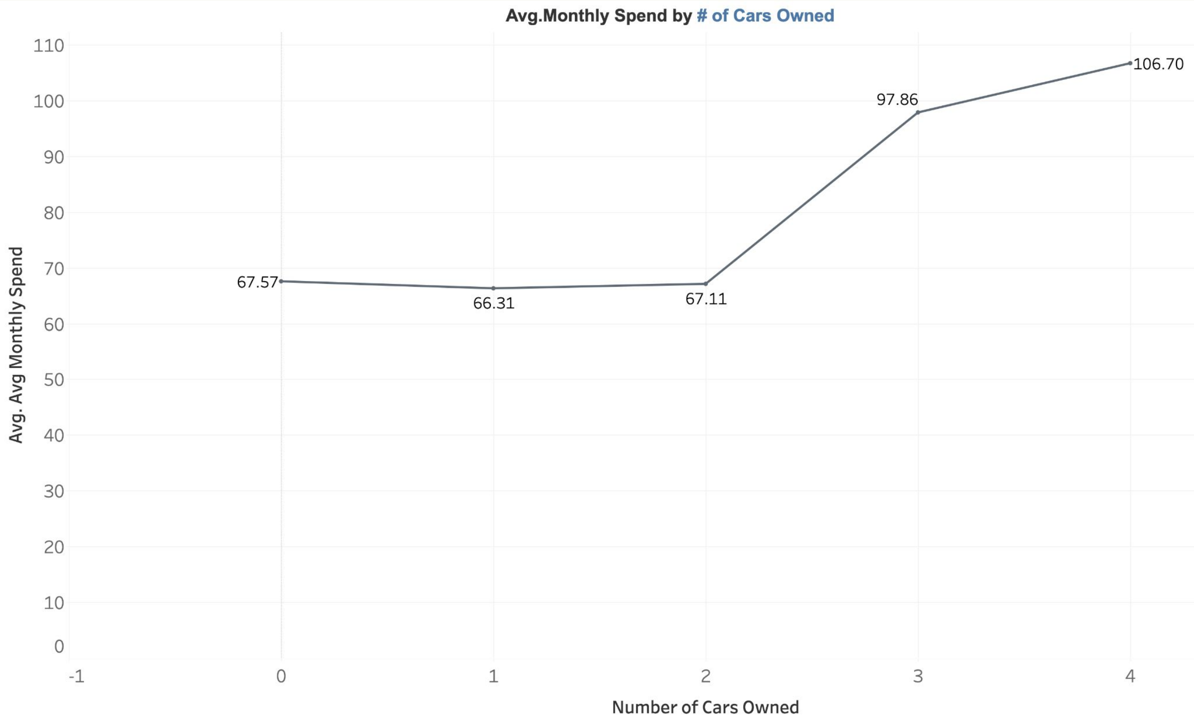


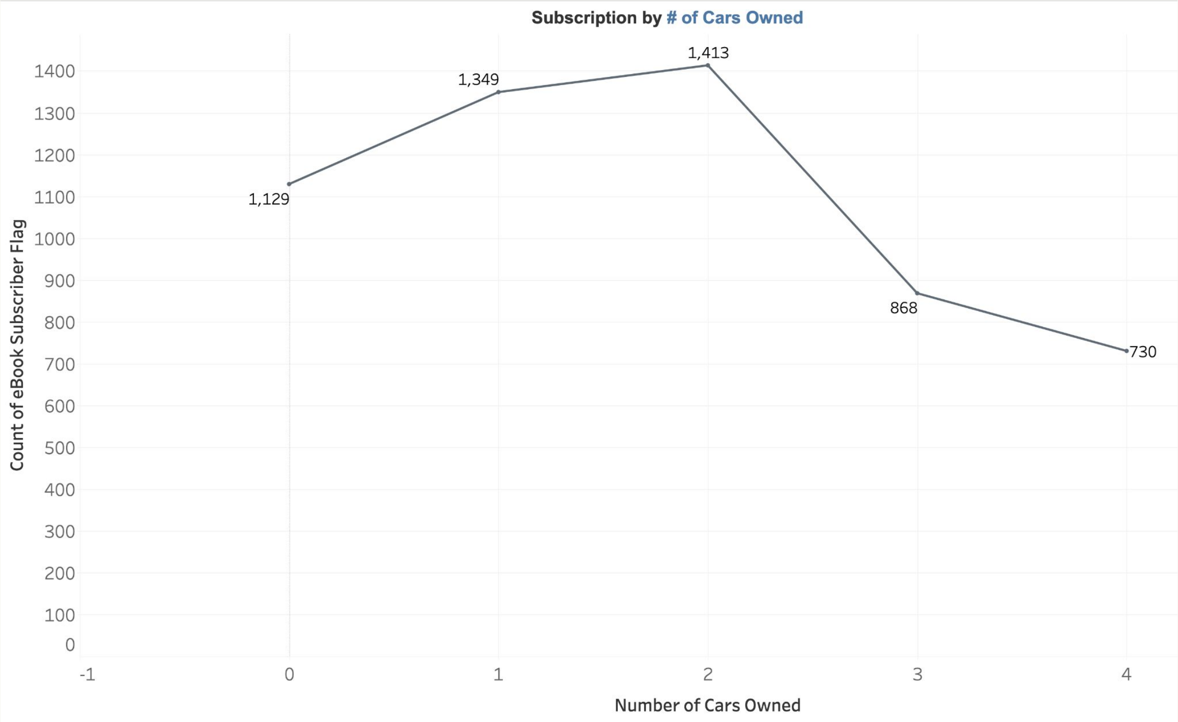


Key Highlights

- ➢ The ***Sweetviz*** report showed insight on specific numbers of the relationship and associations among features.
- ➢ For **Subscription, ‘number of children at home’** had an association of 0.17 and **‘annual income’** had a correlation ratio of 0.29.
- ➢ For **Avg.Monthly Spend, ‘number of children at home’(0.73), ‘gender’(0.57), ‘number of cars owned’(0.47),** and **‘occupation’(0.38)** had a strong correlation ratio.

Data Visualization



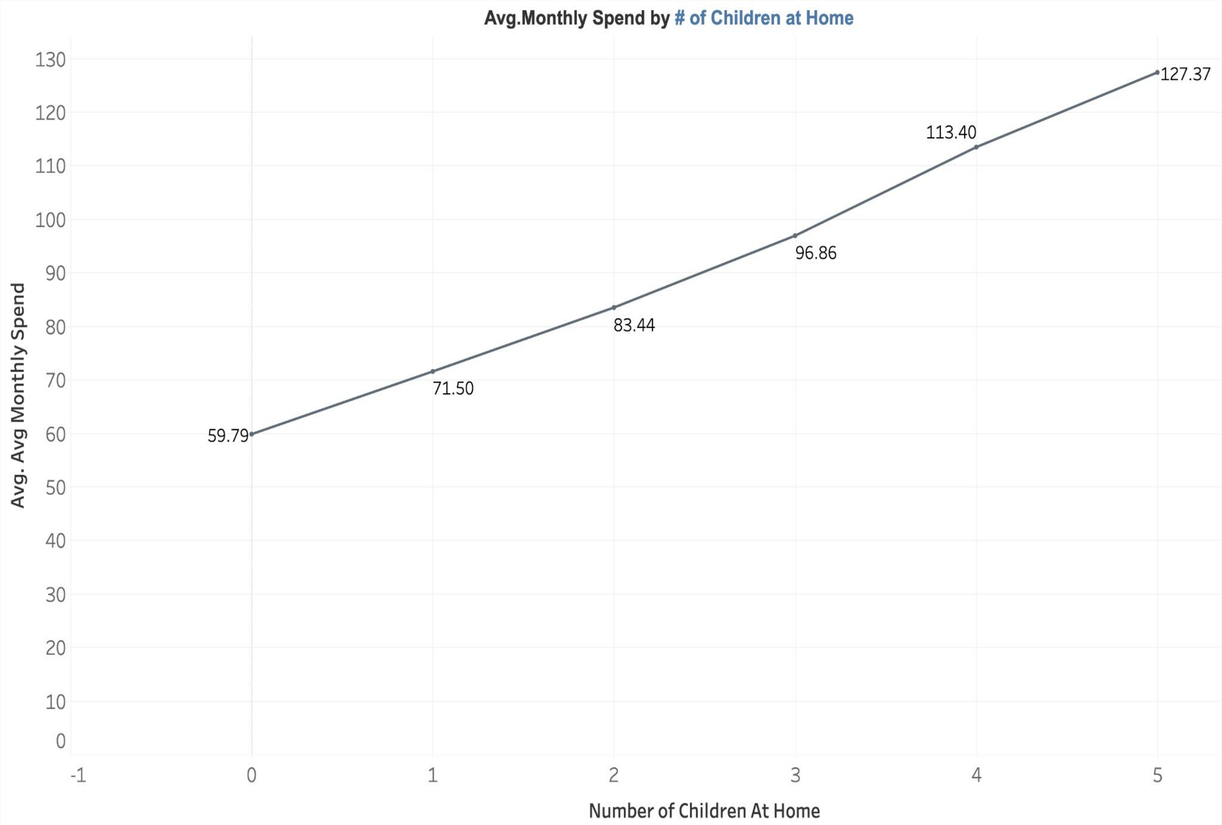


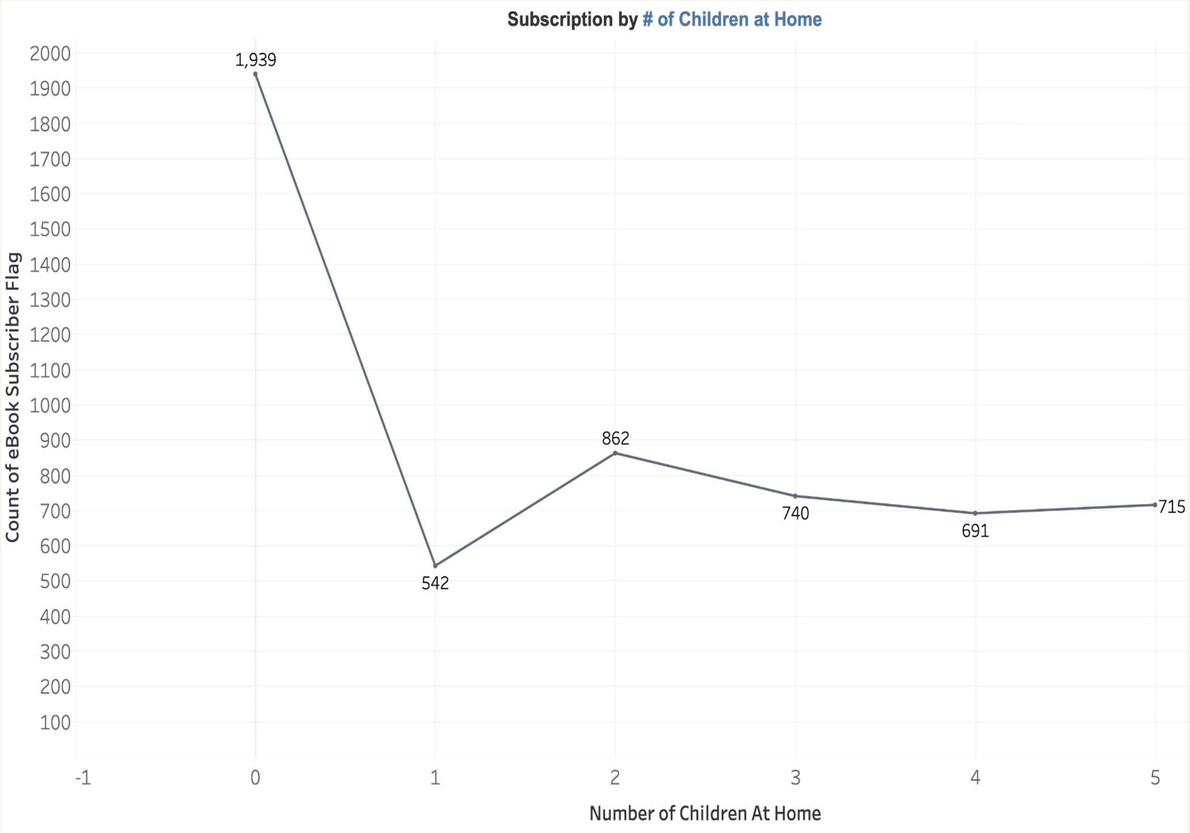
Key Highlights

- ➢ The **number of cars owned** shows an **inverse** relationship between average monthly spend and subscription rates, with **positive correlation with average monthly spend** and a **negative correlation with subscription rates**.
- ➢ Clients with **3 + cars spent significantly more on average**, but those with **0-2 cars had much higher subscription rates**. Subscription numbers drop sharply for clients with 3 or more cars.

–

Data Visualization

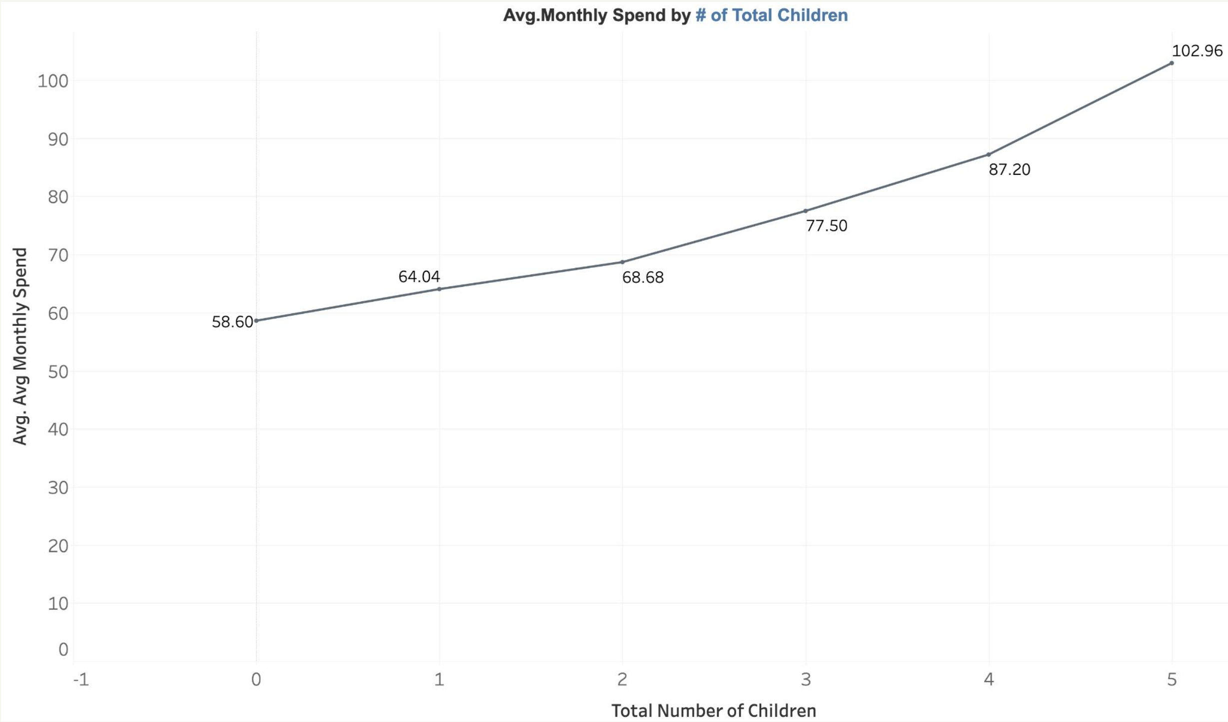


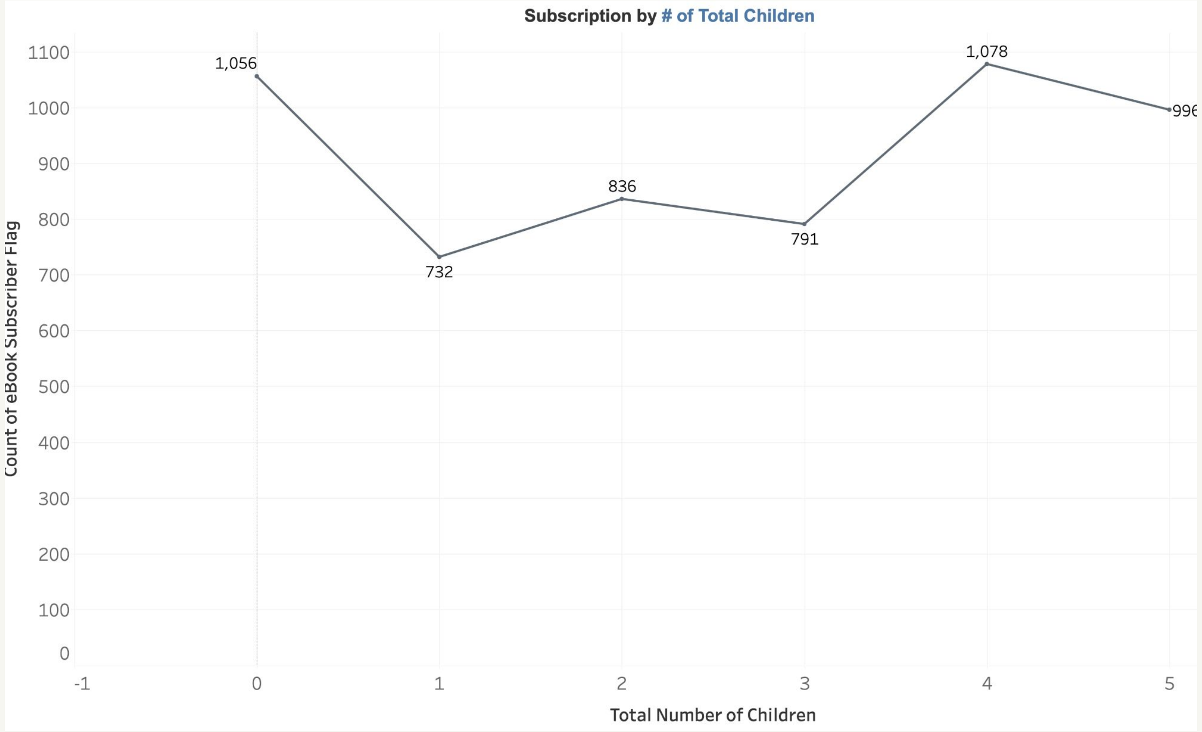


Key Highlights

➢ The **number of children at home** showed a very **strong positive correlation (0.73)** with **average monthly spend**, whereas **clients with no children had the highest subscription rates**.

Data Visualization





Key Highlights

➢ The **total number of children** had a strong positive correlation with **average monthly spend**, but no clear relationship was found with subscription status.

Focused Marketing Strategy for Revenue and Subscriber Growth

✅**Maximize Revenue in High-Potential Markets**

- **Upsell in California:** While California dominates in total clients and spending, the lower average spend per client suggests an opportunity to introduce **premium tiers, exclusive content, and personalized upsell offers** to increase customer value.
- **Expand in Oklahoma:** With high average spend but lower client numbers, **targeted acquisition campaigns** and **localized marketing efforts** can unlock additional revenue potential.

✅ **Leverage Demographic Insights for Targeted Campaigns**

- **Capitalize on High-Spending Males:** Promote **premium offerings, loyalty programs, and exclusive benefits** to males, as they exhibit higher spending and subscription rates.
- **Enhance Engagement with Females:** Develop **personalized promotions** and **tailored content strategies** to increase female spending and subscriptions.
- **Use Education-Based Segmentation:** Prioritize clients with **bachelor's or partial college education** for subscription and upsell campaigns, as they demonstrate stronger engagement.

✅ **Drive Subscription Growth Through Lifestyle & Behavioral Targeting**

- **Target Homeowners for Premium Subscriptions:** Homeowners show stronger financial engagement—offer **high-value subscription packages, extended commitment discounts, or home-related exclusive benefits** to convert them into loyal subscribers.
- **Optimize Offers for Singles & Married Clients:** Singles are more likely to subscribe, making them ideal for **new product launches and promotional offers**. Conversely, encourage higher spending among married clients with **family-centric or bundled subscription options**.
- **Utilize Age-Based Targeting:** Focus on **50s and 60s age groups** as the core subscription base, while implementing engagement campaigns for the **40s and 70s** to increase spending and retention.

Focused Marketing Strategy for Revenue and Subscriber Growth

✅ **Family-Driven Spending Strategy**

- **Monetize Family-Oriented Clients:** Since the n**umber of children** correlates with higher spending, introduce **family plans, parent-focused bundles, and child-friendly add-ons** to further maximize monthly revenue.

✅ **Occupation-Specific Campaigns & Spending Incentives**

- **Convert Managers into Subscribers:** Managers spend more but subscribe less—introduce **career-enhancing content, executive perks, or industry-focused incentives** to encourage subscription.
- **Retain and Reward Professionals:** The professional segment already leads in subscriptions—strengthen loyalty through **tiered rewards, VIP access, and premium engagement opportunities**.

Anomaly Detection and k-means Clustering Model

Pycaret anomalies and clustering model was set up to detect anomalies and cluster the dataset into 4 clusters for the training data

	Education Level	Occupation	Gender	Marital Status	Home Owner Status	Number of Cars Owned	Number of Children At Home	Total Number of Children	Annual Income	Avg Monthly Spend	Age	City	ZipCode	State	Anomaly	Anomaly_Score
0	Bachelors	Professional	M	M	1	0	0	2	137947	89	58	Cleveland	44101	Ohio	0	64.560050
1	Bachelors	Professional	M	S	0	1	3	3	101141	117	59	Seattle	98101	Washington	0	32.756679
2	Bachelors	Professional	M	M	1	1	3	3	91945	123	59	Omaha	68101	Nebraska	0	40.914545
3	Bachelors	Professional	F	S	0	1	0	0	86688	50	56	Fort Worth	76101	Texas	0	37.788887
4	Bachelors	Professional	F	S	1	4	5	5	92771	95	56	Oakland	94601	California	0	38.209946
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16514	Bachelors	Professional	F	M	1	4	5	5	101542	101	59	San Antonio	78201	Texas	0	31.701735
16515	Partial College	Professional	F	S	1	2	0	3	46549	46	88	Pittsburgh	15201	Pennsylvania	0	29.291637
16516	Bachelors	Management	M	M	1	2	0	5	133053	79	85	Honolulu	96801	Hawaii	0	39.786933
16517	High School	Skilled Manual	M	M	1	2	0	4	31930	65	78	Anaheim	92801	California	0	39.912404
16518	High School	Professional	M	S	1	2	0	4	59382	68	79	Fort Wayne	46801	Indiana	0	30.805844

	count
Anomaly	
0	15693
1	826

From the **anomalies** detection, a total of **826** among 16,519 were detected as anomalies and anomaly scores were calculated for the training dataset.

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.5554	55921.6709	0.5250	0	0	0

	count
Cluster	
Cluster 0	5502
Cluster 2	5305
Cluster 1	3402
Cluster 3	2310

From the **clustering** model, the training dataset were divided into 4 clusters.

Monthly Spend Prediction Regression Model Analysis

Pycaret regression model was set up to compare and evaluate all the algorithms to predict the monthly spend.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	2.5106	10.1551	3.1862	0.9863	0.0513	0.0391	2.0450
gbr	Gradient Boosting Regressor	2.5374	10.2048	3.1939	0.9862	0.0508	0.0392	1.7410
xgboost	Extreme Gradient Boosting	2.6473	11.1017	3.3312	0.9850	0.0533	0.0411	0.4750
rf	Random Forest Regressor	2.6705	11.3950	3.3749	0.9846	0.0538	0.0415	6.6350
et	Extra Trees Regressor	2.7694	12.2220	3.4954	0.9835	0.0558	0.0431	5.0080
dt	Decision Tree Regressor	3.5665	20.4585	4.5221	0.9724	0.0723	0.0555	0.2720
lr	Linear Regression	4.7998	40.0598	6.3276	0.9459	0.0903	0.0697	0.7740
ridge	Ridge Regression	4.7990	40.0598	6.3276	0.9459	0.0902	0.0697	0.3450
br	Bayesian Ridge	4.7992	40.0598	6.3276	0.9459	0.0902	0.0697	0.3190
lar	Least Angle Regression	5.0266	43.5348	6.5818	0.9414	0.0958	0.0734	0.2080
ada	AdaBoost Regressor	5.3065	44.0121	6.6307	0.9406	0.0997	0.0814	1.0510
lasso	Lasso Regression	5.1608	50.2794	7.0888	0.9321	0.0873	0.0704	0.2560
llar	Lasso Least Angle Regression	5.1608	50.2789	7.0887	0.9321	0.0873	0.0704	0.2520
en	Elastic Net	10.1146	154.3036	12.4201	0.7917	0.1604	0.1439	0.2600
huber	Huber Regressor	14.8983	393.6696	19.5614	0.4638	0.2479	0.2123	0.2610
omp	Orthogonal Matching Pursuit	16.3105	444.8199	21.0112	0.3982	0.2708	0.2377	0.2250
knn	K Neighbors Regressor	17.5355	524.3236	22.8956	0.2918	0.2944	0.2564	0.2520
dummy	Dummy Regressor	20.5929	742.2433	27.2339	-0.0006	0.3447	0.3057	0.2020
par	Passive Aggressive Regressor	31.9002	1983.4873	38.3466	-1.6620	0.4207	0.4873	0.4080

Among all the evaluated algorithms, '**light gradient boosting machine**' was selected considering it having the lowest overall error including MAE, MSE, RMSE, etc.

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	2.4888	9.9663	3.1569	0.9857	0.0504	0.0385
1	2.5287	10.3550	3.2179	0.9875	0.0515	0.0393
2	2.5338	9.8797	3.1432	0.9872	0.0501	0.0391
3	2.5105	10.5080	3.2416	0.9859	0.0527	0.0398
4	2.4982	9.9136	3.1486	0.9861	0.0514	0.0394
5	2.4200	9.2720	3.0450	0.9875	0.0498	0.0380
6	2.5638	10.3907	3.2235	0.9869	0.0501	0.0392
7	2.5435	10.5924	3.2546	0.9852	0.0518	0.0392
8	2.4907	10.3596	3.2186	0.9854	0.0543	0.0399
9	2.5279	10.3141	3.2116	0.9854	0.0512	0.0390
Mean	2.5106	10.1551	3.1862	0.9863	0.0513	0.0391
Std	0.0378	0.3779	0.0599	0.0009	0.0013	0.0005

Machine Learning model was created using '**lightgbm**' and tested for a 10-fold cross-validation.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	2.3263	8.6922	2.9483	0.9883	0.0480	0.0365

The metrics of the end result was as shown above.

Subscription Prediction Classification Model Analysis

Pycaret classification model was set up to compare and evaluate all the algorithms to predict the likelihood of ebook subscriptions

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.7811	0.8478	0.5908	0.7032	0.6420	0.4861	0.4900	3.5000
dummy	Dummy Classifier	0.6677	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2650

Among all the evaluated algorithms, '**light gradient boosting machine**' was selected considering it having the lowest overall error including Accuracy, AUC, Recall, etc.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7640	0.8296	0.5547	0.6762	0.6094	0.4428	0.4472
1	0.7666	0.8417	0.5688	0.6780	0.6186	0.4524	0.4560
2	0.7891	0.8473	0.5922	0.7238	0.6514	0.5024	0.5076
3	0.7846	0.8591	0.6172	0.6991	0.6556	0.4998	0.5018
4	0.7967	0.8516	0.6120	0.7321	0.6667	0.5221	0.5264
5	0.7803	0.8526	0.5938	0.6994	0.6423	0.4852	0.4886
6	0.7941	0.8651	0.6146	0.7239	0.6648	0.5177	0.5213
7	0.7708	0.8409	0.5938	0.6766	0.6325	0.4669	0.4690
8	0.7993	0.8642	0.6042	0.7436	0.6667	0.5253	0.5311
9	0.7656	0.8259	0.5573	0.6794	0.6123	0.4466	0.4511
Mean	0.7811	0.8478	0.5908	0.7032	0.6420	0.4861	0.4900
Std	0.0129	0.0127	0.0220	0.0245	0.0214	0.0304	0.0306

Machine Learning was created using '**lightgbm**' and tested for a 10-fold cross-validation.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Light Gradient Boosting Machine	0.8486	0.9189	0.6965	0.8207	0.7535	0.6453	0.6498

The metrics of the end result was as shown on the left.

Questions?