5303 Final Project: Food App Business Analysis

Austin Steel

Joshua King

Introduction

The ability of businesses to collect information about their customers allows for businesses to understand their customer base better and keep a high level of engagement with them. This is crucial for business growth in the modern online industry. With the right statistical analysis, key insights can be derived, and actionable steps can be recommended that can make or break a company's comparing customer data and using statistical techniques, we analyzed the dataset titled 'Food App Business' in search of trends and patterns to offer recommendations for growth of the company.

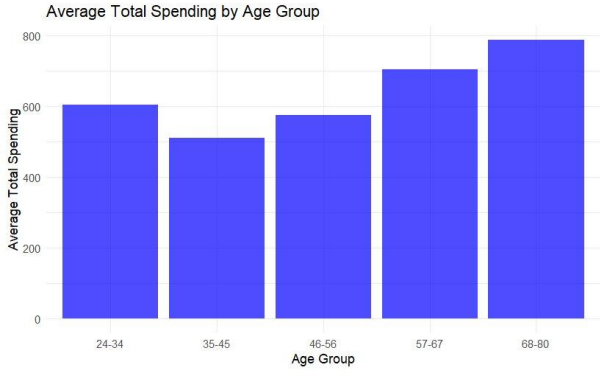
This dataset comes from Kaggle and contains 27 variables on customers purchasing products from either online, instore, or from a catalog. There are 2205 customer entries, where each row is the record for one customer. We are given many different demographic statistics such as age, income, marital status, etc., and how much the customer has spent on various categories such as Wine, Meat, Fruits etc. The purpose of this project is to perform a robust exploratory analysis and produce data driven proposals to add value to the company. Our target variables will be the amount spent on each category and the total amount spent by each customer.

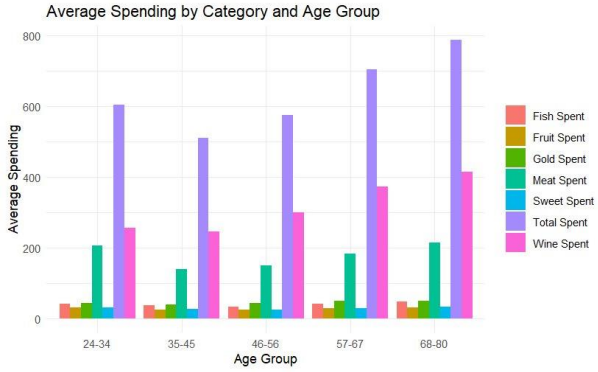
Data Preprocessing / Initial Exploration

To start, we began exploring the data to better understand the information provided. We found that we did not have any missing values, and all entries are integers. Looking at the distribution of each variable gave us some insight into which ones might be the most important. For example, we noticed the *MonthlyIncome* column ranges from $1,730 to $113,734, which is a very large span. Given how much anyone can spend is always related to their income, we believe this variable will be important. The *ActiveSinceDays* column had values all in the 2000s and did not seem to provide any useful information. *Age* spanned from 24 to 80 and will also certainly be another helpful variable. However, typically, companies do not try to target specific ages, but age brackets. Therefore, we added a column titled *Age_Group* which groups customers into 11-year brackets. We combined the *Married* and *Single* columns into a single *Maritial_Status* column since this information was redundant. There are six categories of spending the dataset provides for us. These are: Wines, Fruits, Meat, Fish, Sweet, and Gold. Lastly, we added a *Total_Spent* column, found by summing up all the amounts spent on each category.

```
> summary(df_foodapp)
 MonthlyIncome    ActiveSinceDays      Age          Graduate       NoOfChildren     NoOfTeenager    NoOfDaysSinceLastPurchase
 Min.   :  1730   Min.   :2159    Min.   :24.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.00
 1st Qu.: 35196   1st Qu.:2339    1st Qu.:43.0   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:24.00
 Median : 51287   Median :2515    Median :50.0   Median :1.0000   Median :0.0000   Median :0.0000   Median :49.00
 Mean   : 51622   Mean   :2513    Mean   :51.1   Mean   :0.8857   Mean   :0.2921   Mean   :0.3351   Mean   :49.01
 3rd Qu.: 68281   3rd Qu.:2688    3rd Qu.:61.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:74.00
 Max.   :113734   Max.   :2858    Max.   :80.0   Max.   :1.0000   Max.   :2.0000   Max.   :2.0000   Max.   :99.00
 AmountSpendOnWines AmountSpentOnFruits AmountSpentOnMeat AmountSpentOnFish AmountSpentOnSweet AmountSpentOnGold
 Min.   :   0.0     Min.   :  0.0       Min.   :   0.0    Min.   :  0.00    Min.   :  0.00     Min.   :  0.00
 1st Qu.:  24.0     1st Qu.:  2.0       1st Qu.:  16.0    1st Qu.:  3.00    1st Qu.:  1.00     1st Qu.:  9.00
 Median : 178.0     Median :  8.0       Median :  68.0    Median : 12.00    Median :  8.00     Median : 25.00
 Mean   : 306.2     Mean   : 26.4       Mean   : 165.3    Mean   : 37.76    Mean   : 27.13     Mean   : 44.06
 3rd Qu.: 507.0     3rd Qu.: 33.0       3rd Qu.: 232.0    3rd Qu.: 50.00    3rd Qu.: 34.00     3rd Qu.: 56.00
 Max.   :1493.0     Max.   :199.0       Max.   :1725.0    Max.   :259.00    Max.   :262.00     Max.   :321.00
 NoOfDealsWithDiscount NoOfWebPurchase  NoOfCatalogPurchase NoOfStorePurchase NoOfWebVisitsMonth PurchasedIn1stCampaign
 Min.   : 0.000        Min.   : 0.000   Min.   : 0.000      Min.   : 0.000    Min.   : 0.000     Min.   :0.0000
 1st Qu.: 1.000        1st Qu.: 2.000   1st Qu.: 0.000      1st Qu.: 3.000    1st Qu.: 3.000     1st Qu.:0.0000
 Median : 2.000        Median : 4.000   Median : 2.000      Median : 5.000    Median : 6.000     Median :0.0000
 Mean   : 2.318        Mean   : 4.101   Mean   : 2.645      Mean   : 5.824    Mean   : 5.337     Mean   :0.0644
 3rd Qu.: 3.000        3rd Qu.: 6.000   3rd Qu.: 4.000      3rd Qu.: 8.000    3rd Qu.: 7.000     3rd Qu.:0.0000
 Max.   :15.000        Max.   :27.000   Max.   :28.000      Max.   :13.000    Max.   :20.000     Max.   :1.0000
 PurchasedIn2ndCampaign PurchasedIn3rdCampaign PurchasedIn4thCampaign PurchasedIn5thCampaign TotalNoOfCampaignAccepted
 Min.   :0.00000        Min.   :0.00000        Min.   :0.00000        Min.   :0.00000        Min.   :0.0000
 1st Qu.:0.00000        1st Qu.:0.00000        1st Qu.:0.00000        1st Qu.:0.00000        1st Qu.:0.0000
 Median :0.00000        Median :0.00000        Median :0.00000        Median :0.00000        Median :0.0000
 Mean   :0.01361        Mean   :0.07392        Mean   :0.07438        Mean   :0.07302        Mean   :0.2993
 3rd Qu.:0.00000        3rd Qu.:0.00000        3rd Qu.:0.00000        3rd Qu.:0.00000        3rd Qu.:0.0000
 Max.   :1.00000        Max.   :1.00000        Max.   :1.00000        Max.   :1.00000        Max.   :4.0000
 CustomerComplain  Marital_Status      Total_Spent     Age_Group
 Min.   :0.00000   Length:2205        Min.   :   5.0   24-34:219
 1st Qu.:0.00000   Class :character   1st Qu.:  69.0   35-45:617
 Median :0.00000   Mode  :character   Median : 397.0   46-56:683
 Mean   :0.00907                      Mean   : 606.8   57-67:510
 3rd Qu.:0.00000                      3rd Qu.:1047.0   68-80:176
 Max.   :1.00000                      Max.   :2525.0
```

To finish our exploratory analysis, we wanted to create some basic visualizations of spending. The first graph is the average total spending by age group.



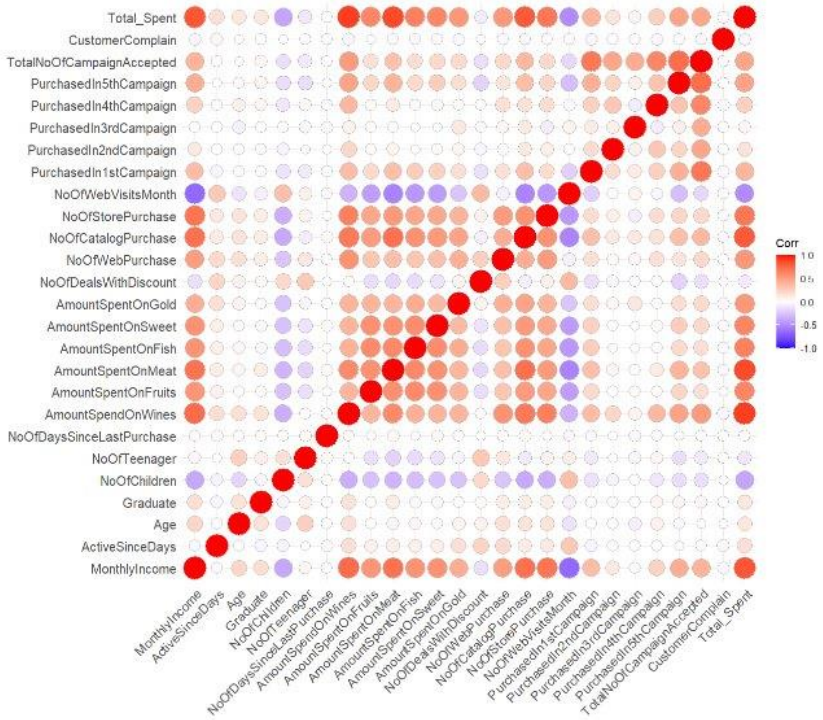


Here, we can see that the oldest age groups spend the most, and the 35 – 45 age group spends the least. In the second graph, we look at average spending by category among the age groups.

The first thing to note here is that every age group spends the most on Wine first and Meat second. These two categories far outweigh the other four categories tracked. Based on these two graphs we can tell which age groups spend the most/least and on which categories. With this information, we would recommend the business target the lowest spending age group (35 – 45) with advertisements or discounts to try to increase customer spending. Also, since the customers were buying Wine and Meat, but not as much of the other products, the business could offer combination deals to try to increase sales in those

lower categories. For example, the company could offer a discount for Sweets after the first $100 dollars spent on Wine. Since customers are already spending more than $100 on wine, this could entice them to additionally spend on Sweets where they otherwise would not.

To finish our initial data exploration, we created a correlation matrix heatmap to visualize the relationship between the variables. The matrix allows us to compare the correlation of each numerical variable. The deeper a color, the stronger the correlation, with red being positive and blue being negative. The correlation represents the strength of the relationship between two variables. MonthlyIncome was seen to have some strong or moderate correlation with many variables. We also saw that there was some moderate correlation related to the spending on the different food
categories.



Modeling and Testing

From here, we conducted an ANOVA and TUKEY test to compare the amount spent on each category and total by age group. This gave us some hard numbers for these differences. Below are some of the results.

Results:

Total Spent:

68-80 age group spent more than 24-34 group by $184.34

68-80 age group spent more than 35-45 group by $277.51

68-80 age group spent more than 45-56 group by $214.10

57-67 age group spent more than 35-45 group by $192.40

57-67 age group spent more than 46-56 group by $128.99

Wine:

68-80 age group spent more than 24-34 group by $158.10

68-80 age group spent more than 35-45 group by $168.24

68-80 age group spent more than 45-56 group by $114.69

57-67 age group spent more than 24-34 group by $116.99

57-67 age group spent more than 35-45 group by $127.13

46-56 age group spent more than 35-45 group by $53.56

Meat:

68-80 age group spent more than 35-45 group by $74.34

68-80 age group spent more than 46-56 group by $63.56

57-67 age group spent more than 35-45 group by $44.64

24-34 age group spent less than 35-45 group by $66.22

24-34 age group spent more than 46-56 group by $55.44

Fish:

68-80 age group spent more than 46-56 group by $15.17

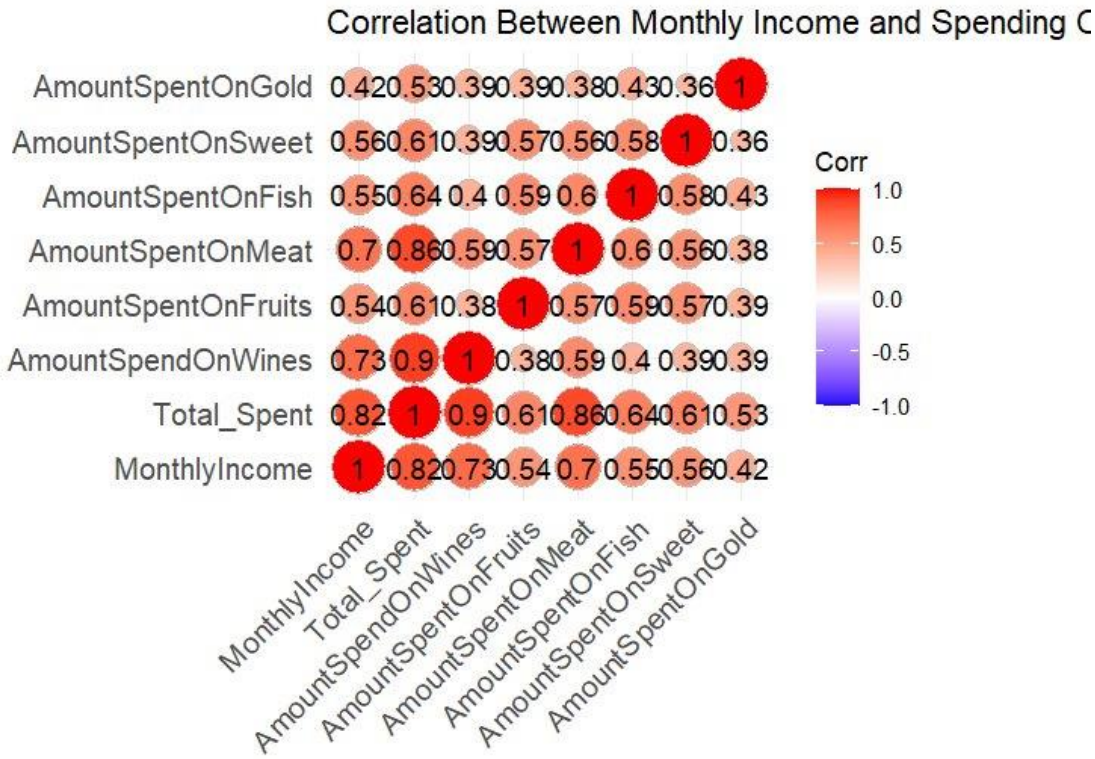
57-67 age group spent more than 46-56 group by $8.80

Gold:

57-67 age group spent more than 35-45 group by $8.97

Again, we see the older age groups are spending significantly more than the younger age groups. There could be many reasons for this, one may simply be that older customers have more income or money to spend. From here, we thought it would be helpful to look

specifically at the correlations between spending on each category and total spending, to Monthly Income. We created a correlation heatmap to visualize this below. Immediately we noticed we have high correlations between Total_Spent, Amount Spent on Wines, and Amount Spent on Meats with Monthly Income. We see moderate correlations with Amont Spent on Fish, Amount Spent on Fruits, and Amount Spent on Sweets and Monthly Income. And we have a weak correlation with the Amount Spent on Gold and Monthly Income. This matches with our initial thoughts that Monthly Income would be a large factor in the Amount Spent on anything, given customers can only spend the money they have.



Since we have several strong correlations between categories of spending and Monthly Income, we decided to build a linear regression model for each category. Monthly Income is the input (x) for each formula and amount spent (total or categorical) is the output (y). For each model, our P-values are all less than 0.05, and therefore statistically significant. We produced seven linear equations which can predict how much a customer spends for each category and the total, based off their standardized Monthly Income.

Total Spending: $y = 606.82 + 495.64x$

Wine: $y = 306.17 + 246.54x$

Fruit: $y = 26.40 + 21.40x$

Meat: $y = 165.31 + 152.99x$

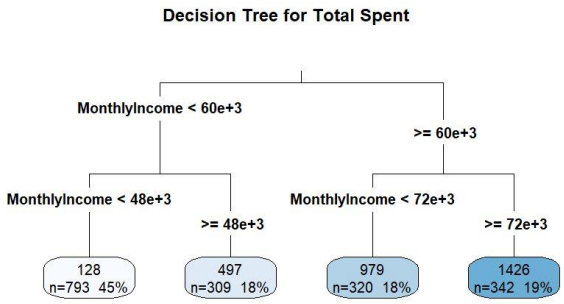
Fish: y = 37.76 + 30.25x

Sweets: y = 27.13 + 22.85x

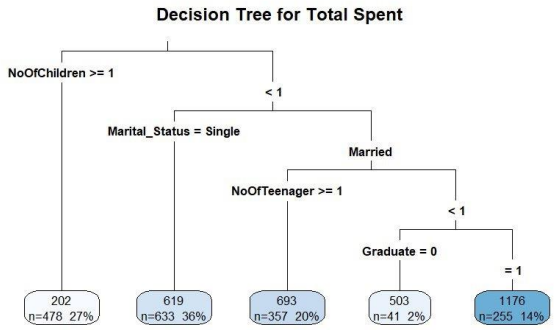
Gold: y = 44.06 + 21.61x

These linear relationships can be used to help assess customer spending. Given a customer's standardized Monthly Income, we can use these regression models to predict how much we would expect them to spend on each category. Customers spending more than this can be offered rewards to keep their engagement high, and customers spending less than expected can be targeted with advertisements or discounts to increase their engagement. Customers spending significantly more or less than expected should be flagged and studied further to gain valuable insights.

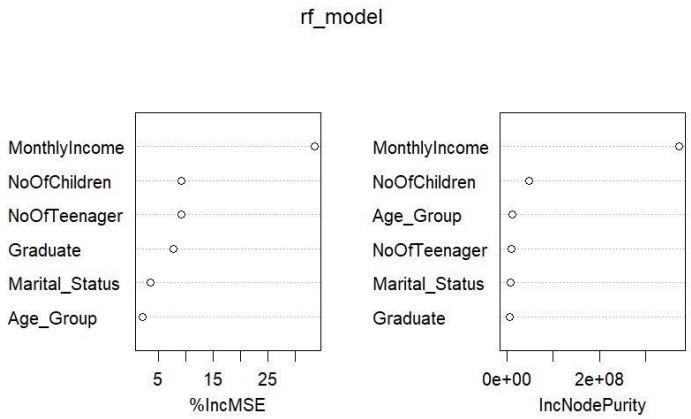
At this stage, we decided to create a decision tree model for Total_Spent. The first model created just broke down customers into categories based on Monthly Income. Therefore, we created a second model specifically looking at the other



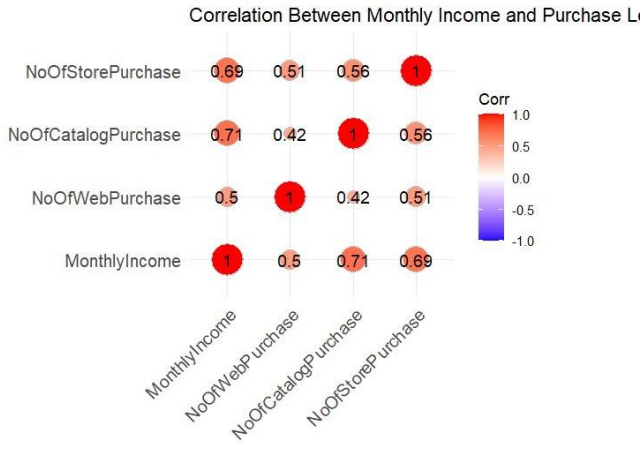
variables.



We also created a random forest model to visualize the impact of each variable. We can see from this model that Monthly Income truly is the strongest indicator of Total Spent, with the second most important factor being Number of Children. Graduate or not, Marital Status, and Age Group are the next most important factors in Total Spent after that.



Lastly, we thought we would check the correlations between Monthly Income and Purchase Locations. Our dataset gives us the number of purchases that are Store, Catalog, and Web for each customer. When we looked at the correlations here, we see Monthly Income is moderately to strongly correlated with all three.



Since we do see some strong correlations here, we decided to build a linear regression model for each of these variables as well with Monthly Income. These models give us predictions as to how many purchases a customer should have in Store, Catalog, and Web based off their Monthly Income. Our P-value for each model is much less than 0.05.

- Web: $y = 0.6678 + 0.0000665x$
- Catalog: $y = -2.307 + 0.00009594x$
- Store: $y = 0.2714 + 0.0001076x$

We can again use these regression models just like our previous ones to compare actual customer purchases to expected customer purchases. Customers that are purchasing significantly more or less than predicted should be flagged and looked at further.

Conclusions

After our analysis, we reached a few conclusions.

1. Monthly Income is the most key factor in how much a customer spends.
2. The oldest age groups (57 – 67, and 68 – 80) spent the most in all food categories.
3. For all age groups, we see that customers spend most on Wine, and second most on Meat.

From these findings, we can make many different recommendations to the company to help business growth. For example, we know the 35 – 45 age group spends the least overall and it would be worth targeting with ad campaigns to try to increase spending. Since all customers appear to be spending a lot on Wine and Meat, we would recommend creating deals to pair other products to these two. With our linear regression models, we can predict how much a customer should spend in each category and total from their Monthly Income. This can be helpful in identifying customers who are underspending for their Monthly Income and may have more disposable income they are willing to spend on the company.

References

Dataset: https://www.kaggle.com/datasets/ybifoundation/food-app-business?resource=download