# Consumer Profiles: Prediction

Gabe Anoia, Harrison Hubbard,
Levi Sessions

# Customer Personality Analysis Dataset

https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

- 2,450 samples over 28 features
- Customer demographic, past spending, responses to marketing
- Data collected from undisclosed European retail company
- Collected from sales data and customer survey responses

# Demographic Features

| | | |
|---|---|---|
| **Year_Birth** | Customer's Birth Year | Categorical |
| **Education** | Customer's education level | Categorical |
| **Marital_Status** | Customer's marital status | Categorical |
| **Income** | Customer's yearly household income | Numerical |
| **Kidhome** | Number of children in customer's household | Numerical |
| **Teenhome** | Number of teenagers in customer's household | Numerical |
| **Dt_Customer** | Date of customer's enrollment with the company | Categorical |
| **Recency** | Number of days since customer's last purchase | Numerical |

# Behavioral Features



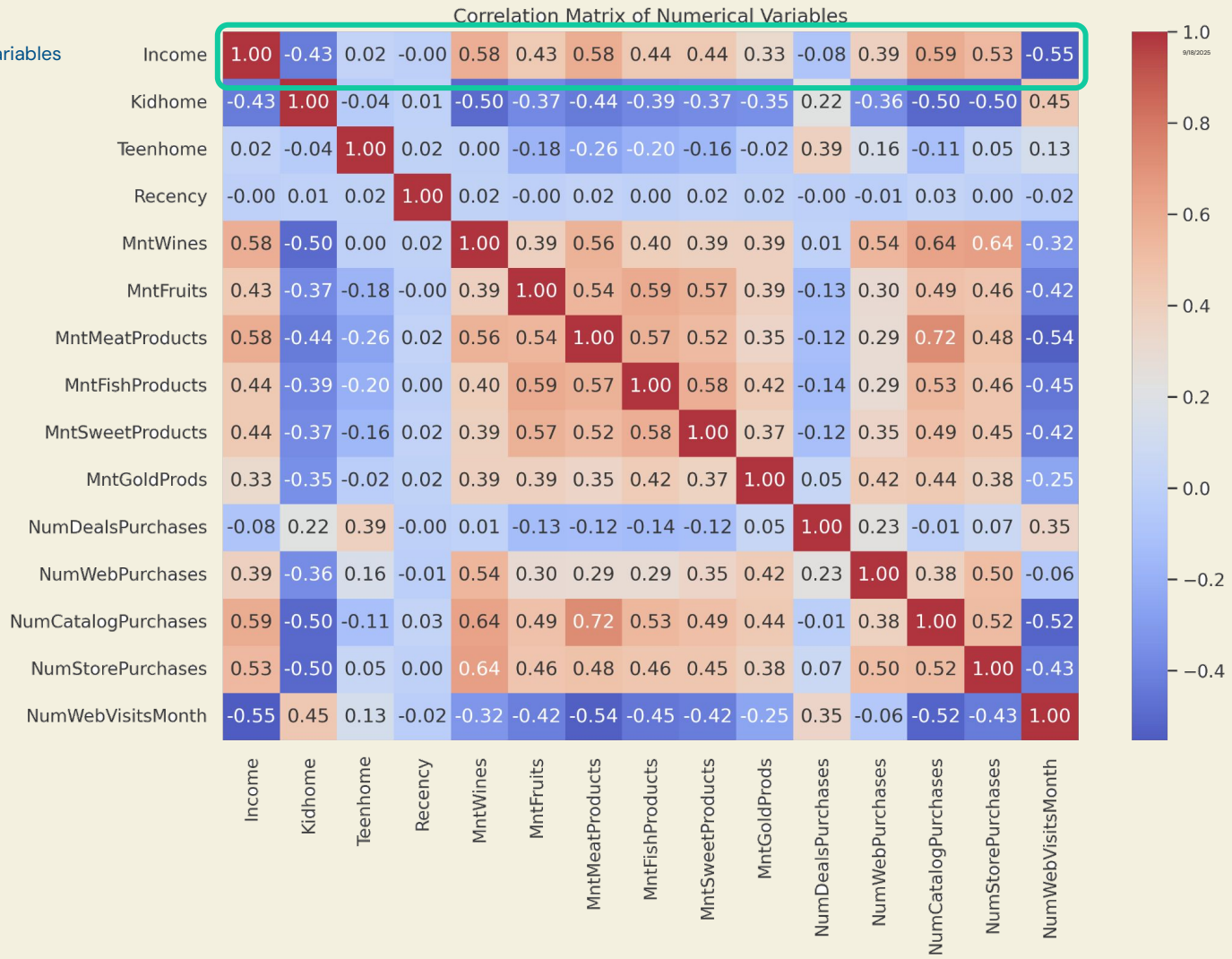
| | | |
|---|---|---|
| **MntWines** | Amount spent on wine in last 2 years | Numerical |
| **MntFruits** | Amount spent on fruits in last 2 years | Numerical |
| **MntMeatProducts** | Amount spent on meat in last 2 years | Numerical |
| **MntFishProducts** | Amount spent on fish in last 2 years | Numerical |
| **MntSweetProducts** | Amount spent on sweets in last 2 years | Numerical |
| **MntGoldProds** | Amount spent on gold in last 2 years | Numerical |
| **NumWebPurchases** | Number of purchases made through the company's website | Numerical |
| **NumCatalogPurchases** | Number of purchases made using a catalogue | Numerical |
| **NumStorePurchases** | Number of purchases made directly in stores | Numerical |
| **NumWebVisitsMonth** | Number of visits to company's website in the last month | Numerical |

# Engagement Features

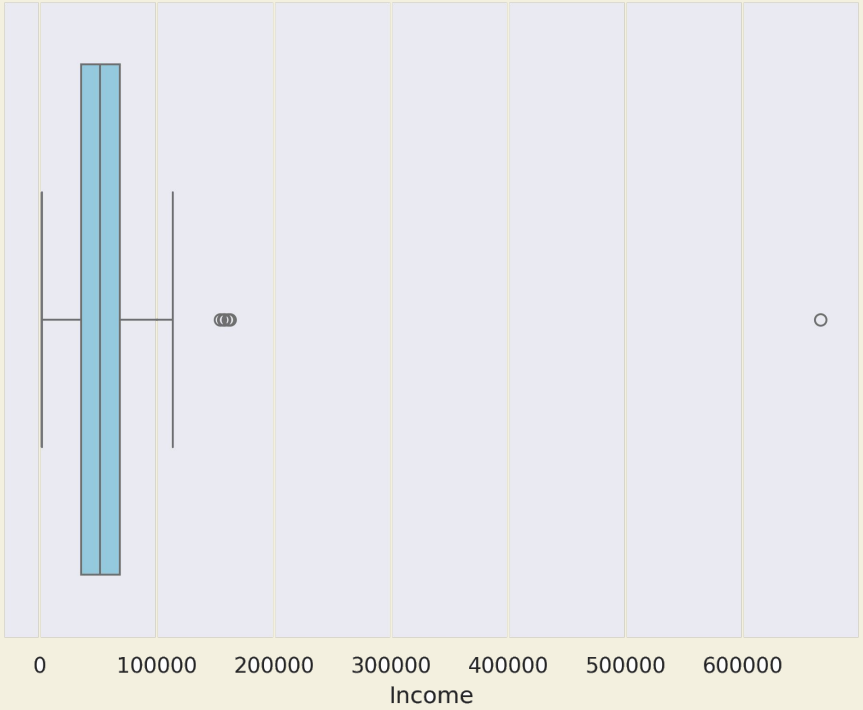
| | | |
|---|---|---|
| **NumDealsPurchases** | Number of purchases made with a discount | Categorical |
| **AcceptedCmp1** | If customer accepted the offer in the 1st campaign | Categorical |
| **AcceptedCmp2** | If customer accepted the offer in the 2nd campaign | Categorical |
| **AcceptedCmp3** | If customer accepted the offer in the 3rd campaign | Categorical |
| **AcceptedCmp4** | If customer accepted the offer in the 4th campaign | Categorical |
| **AcceptedCmp5** | If customer accepted the offer in the 5th campaign | Categorical |
| **Response** | If customer accepted the offer in the last campaign | Categorical |

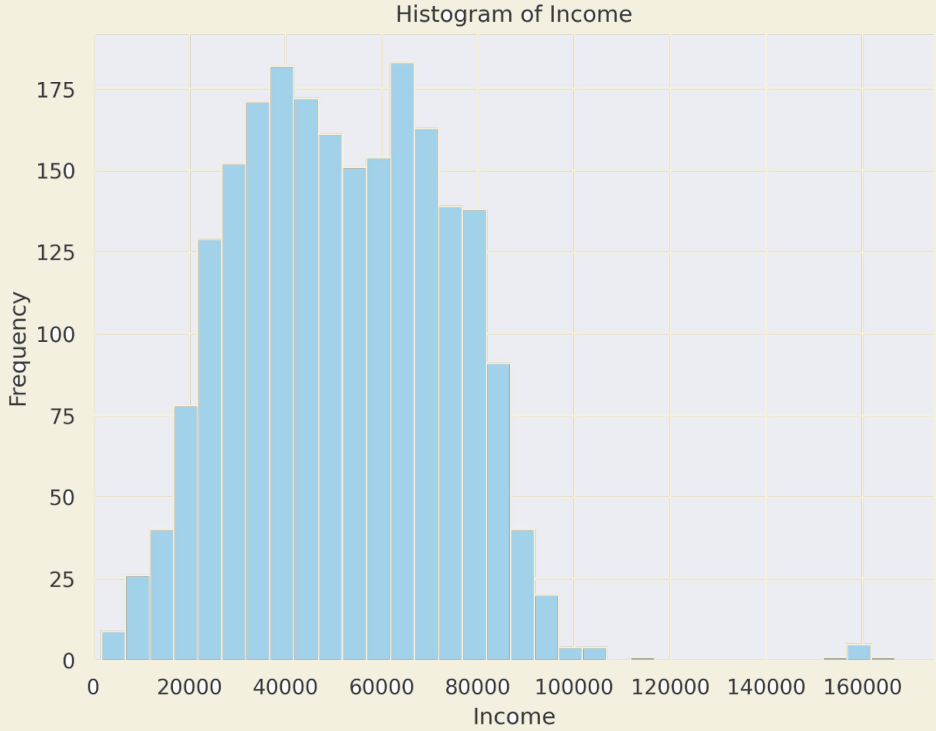
# Our Goals

- Use linear regression to find model used to predict income based on subset of factors with high predictive power
- Use a variety of classification techniques to predict binary income categories with a high degree of accuracy

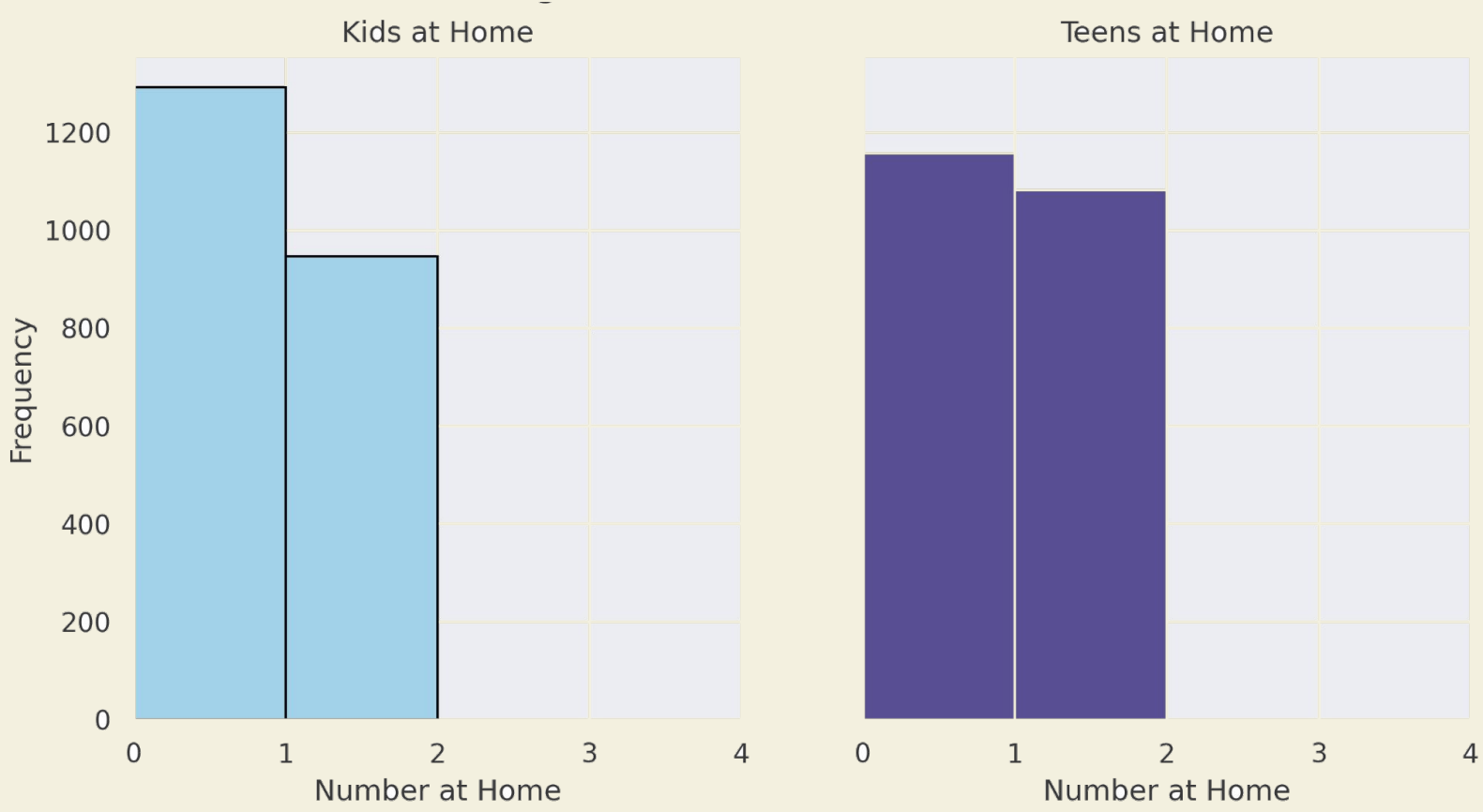

Correlation Matrix of Numerical Variables
Income 1.00 -0.43 0.02 -0.00 0.58 0.43 0.58 0.44 0.44 0.33 -0.08 0.39 0.59 0.53 -0.55
Kidhome -0.43 1.00 -0.04 0.01 -0.50 -0.37 -0.44 -0.39 -0.37 -0.35 0.22 -0.36 -0.50 -0.50 0.45
Teenhome 0.02 -0.04 1.00 0.02 0.00 -0.18 -0.26 -0.20 -0.16 -0.02 0.39 0.16 -0.11 0.05 0.13
Recency -0.00 0.01 0.02 1.00 0.02 -0.00 0.02 0.00 0.02 0.02 -0.00 -0.01 0.03 0.00 -0.02
MntWines 0.58 -0.50 0.00 0.02 1.00 0.39 0.56 0.40 0.39 0.39 0.01 0.54 0.64 0.64 -0.32
MntFruits 0.43 -0.37 -0.18 -0.00 0.39 1.00 0.54 0.59 0.57 0.39 -0.13 0.30 0.49 0.46 -0.42
MntMeatProducts 0.58 -0.44 -0.26 0.02 0.56 0.54 1.00 0.57 0.52 0.35 -0.12 0.29 0.72 0.48 -0.54
MntFishProducts 0.44 -0.39 -0.20 0.00 0.40 0.59 0.57 1.00 0.58 0.42 -0.14 0.29 0.53 0.46 -0.45
MntSweetProducts 0.44 -0.37 -0.16 0.02 0.39 0.57 0.52 0.58 1.00 0.37 -0.12 0.35 0.49 0.45 -0.42
MntGoldProds 0.33 -0.35 -0.02 0.02 0.39 0.39 0.35 0.42 0.37 1.00 0.05 0.42 0.44 0.38 -0.25
NumDealsPurchases -0.08 0.22 0.39 -0.00 0.01 -0.13 -0.12 -0.14 -0.12 0.05 1.00 0.23 -0.01 0.07 0.35
NumWebPurchases 0.39 -0.36 0.16 -0.01 0.54 0.30 0.29 0.29 0.35 0.42 0.23 1.00 0.38 0.50 -0.06
NumCatalogPurchases 0.59 -0.50 -0.11 0.03 0.64 0.49 0.72 0.53 0.49 0.44 -0.01 0.38 1.00 0.52 -0.52
NumStorePurchases 0.53 -0.50 0.05 0.00 0.64 0.46 0.48 0.46 0.45 0.38 0.07 0.50 0.52 1.00 -0.43
NumWebVisitsMonth -0.55 0.45 0.13 -0.02 -0.32 -0.42 -0.54 -0.45 -0.42 -0.25 0.35 -0.06 -0.52 -0.43 1.00
Income
Kidhome
Teenhome
Recency
MntWines
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
MntGoldProds
NumDealsPurchases
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

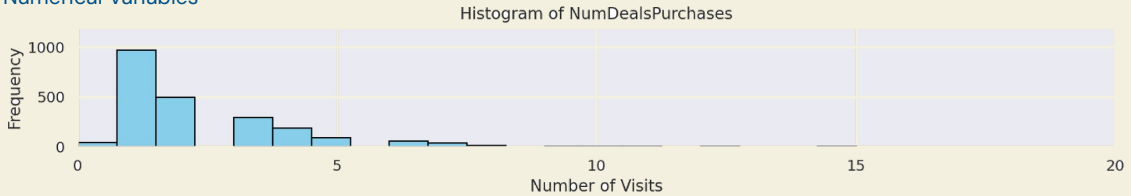
# Income

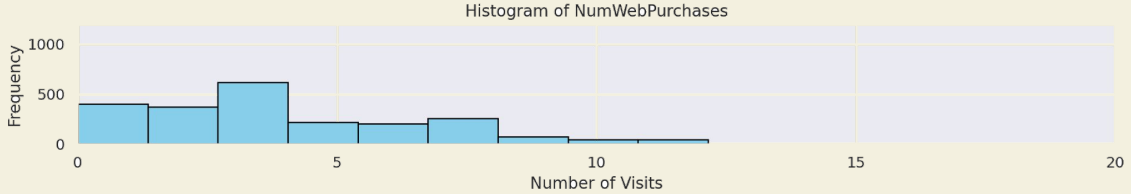

0
100000
200000
300000
400000
500000
600000
Income

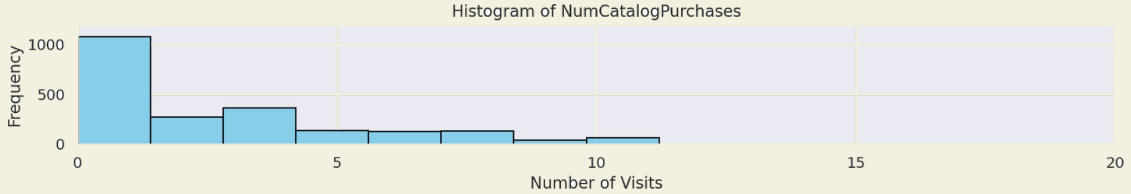

Histogram of Income
Frequency
0
25
50
75
100
125
150
175
0
20000
40000
60000
80000
100000
120000
140000
160000
Income

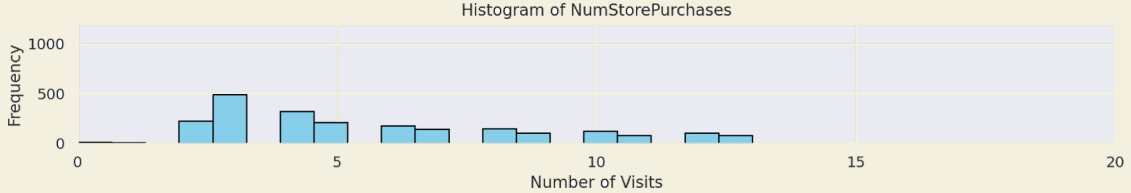
# Kids and Teens in Home

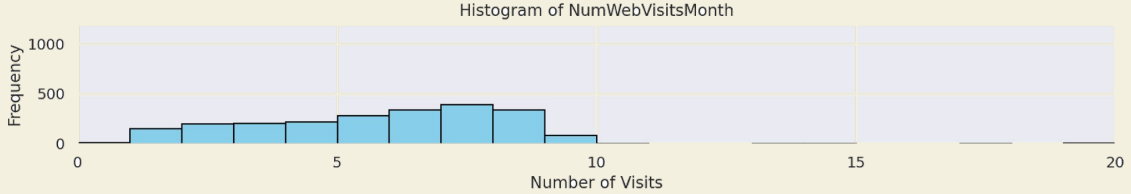




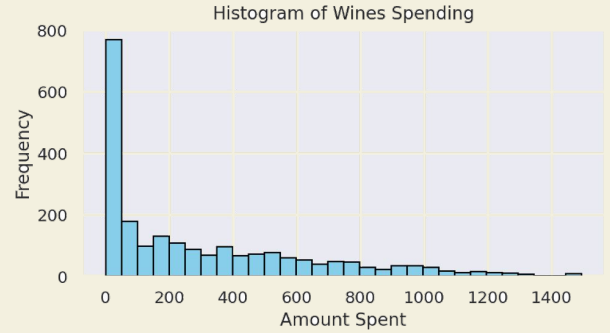



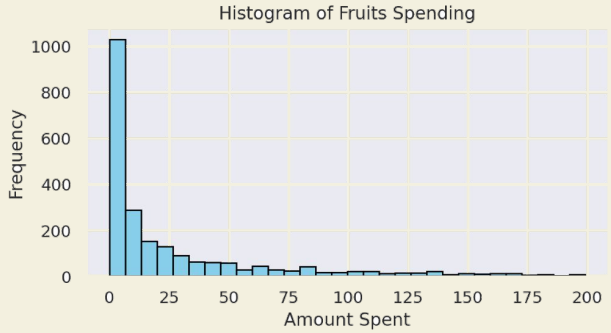

## Notable Conclusions:

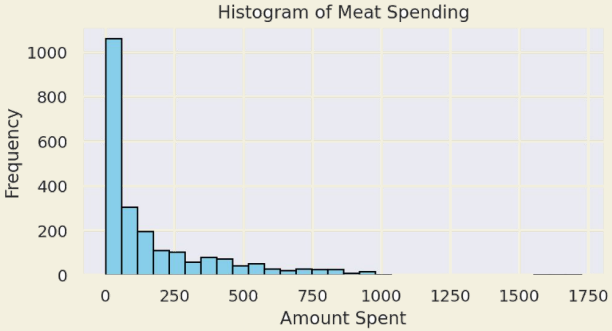
- The distributions of purchases with deals, from catalogs, and from the store are all slightly skewed to the right
- The distribution of the web purchases made is roughly symmetrical, while the number of web visits per month has a slight left skew

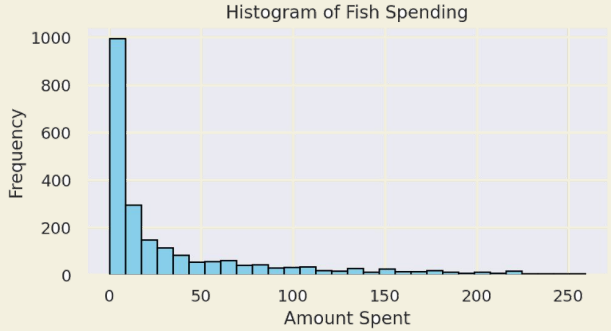
# Notable Conclusions:

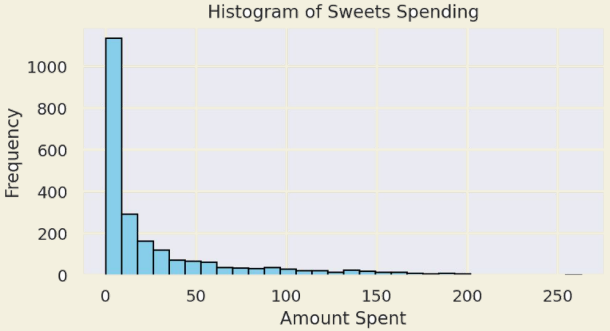
- Hard right skew in all distributions of the spending types
- Wine has the largest spread of the items, while also having the longest right tail of them all
- Gold and meat have a large cluster of data within the first $200 of spending

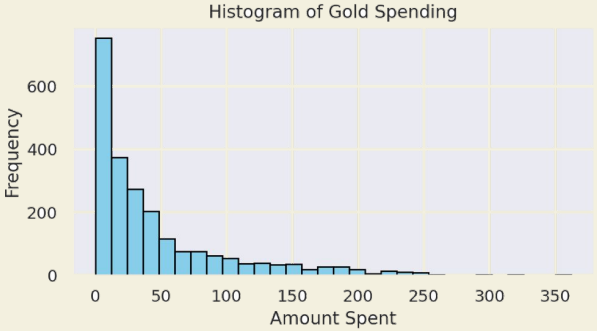


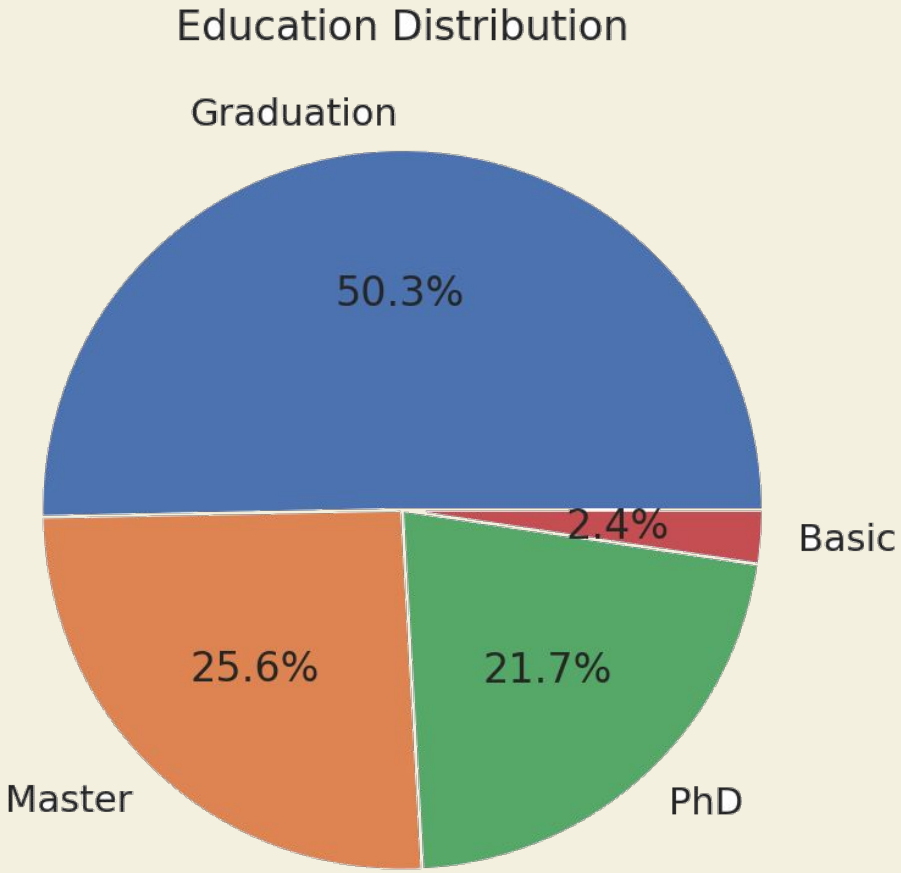



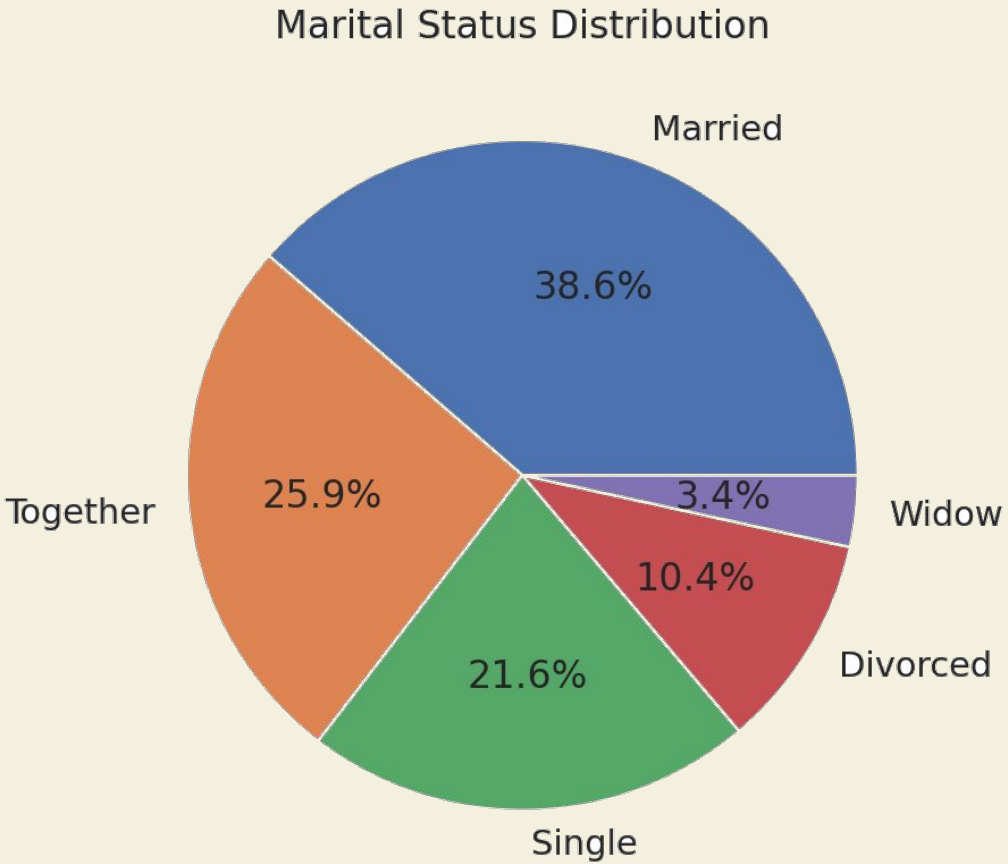






## Notable Conclusions:

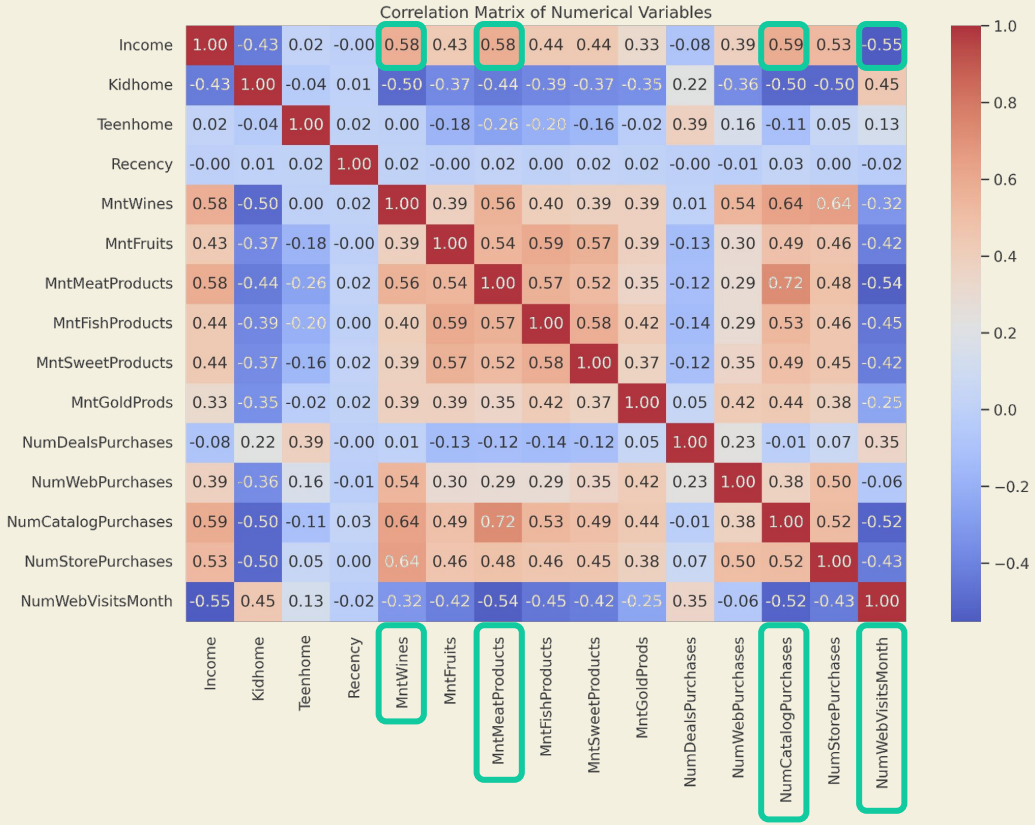
- Majority of customers are highly educated (Graduation, Master's, PhD), with only a small fraction reporting basic education.
- Higher education levels suggest a customer base with greater income potential and stronger spending in premium categories like wine and meat.




## Notable Conclusions:

- Multiple categories representing similar structures => Combine into Married, Divorced, and Single

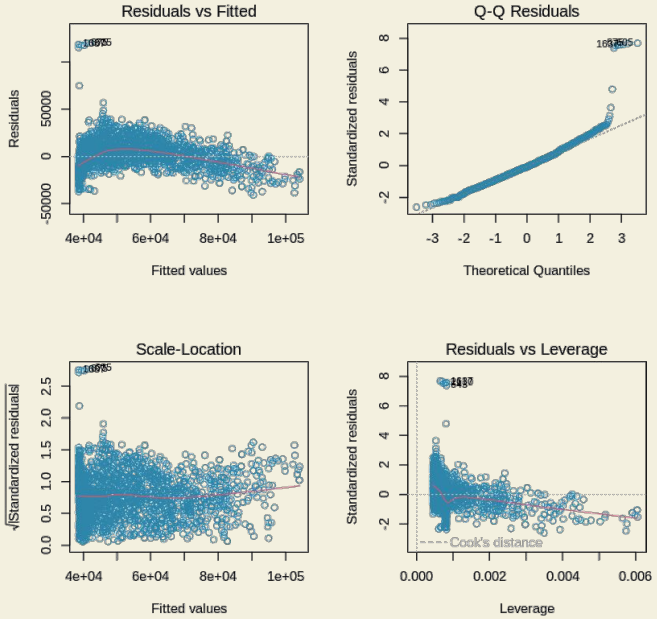
- Household structure likely influences spending priorities => Families with children showed lower non essential spending, while singles may focus more on time saving purchases.




# Choosing Covariates

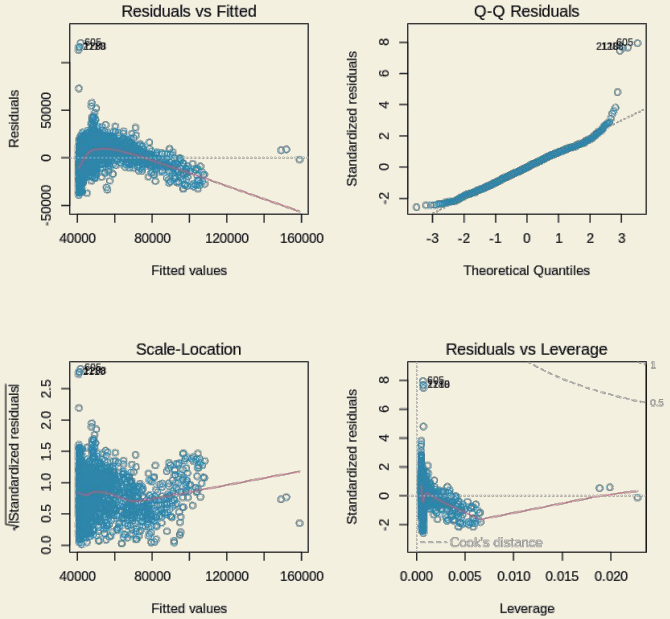
- Pick four variables with highest correlation
- MntWines
- MntMeatProducts
- NumCatalogPurchases
- NumWebVisitsMonth

## Income ~ MntWines



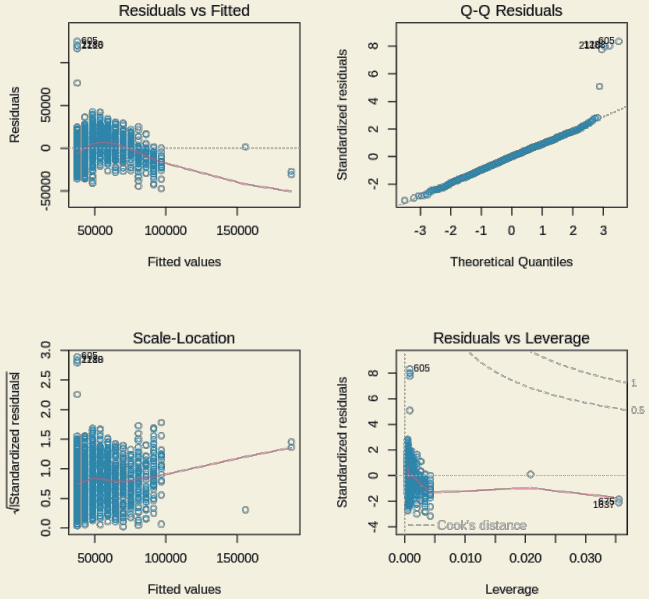

| | |
|---|---|
| Intercept: | **38,600** |
| Coefficient: | **43.88** |
| $R^2$: | **0.4738** |

## Income ~ MntMeatProducts



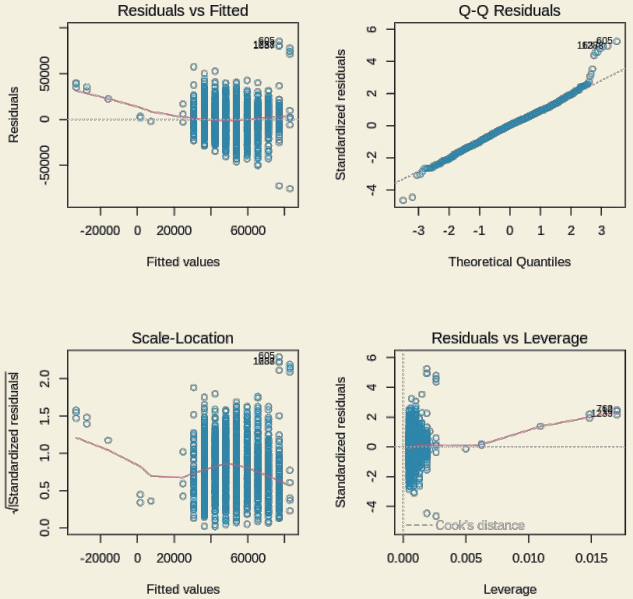

| | |
|---|---|
| Intercept: | **40,567.33** |
| Coefficient: | **68.65** |
| $R^2$: | **0.5017** |

## Income ~ NumCatalogPurchases




| | |
|---|---|
| Intercept: | **37,698.30** |
| Coefficient: | **5,371.10** |
| $R^2$: | **0.5165** |

## Income ~ NumWebVisitMonth




| | |
|---|---|
| Intercept: | **82,855.00** |
| Coefficient: | **-5,802.60** |
| $R^2$: | **0.4275** |

# “Base” Model

```
Residuals:
   Min     1Q Median     3Q    Max
-76638  -5805   -135   5477 108928

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            50256.450   7376.772   6.813 1.23e-11 ***
EducationBasic        -10448.429   1581.847  -6.605 4.97e-11 ***
EducationGraduation       758.418    786.927   0.964 0.335267
EducationMaster          1255.406    913.429   1.374 0.169462
EducationPhD             2146.400    893.553   2.402 0.016384 *
Marital_StatusDivorced  -2888.356   7300.836  -0.396 0.692424
Marital_StatusMarried   -3433.096   7274.632  -0.472 0.637026
Marital_StatusSingle    -3840.489   7280.916  -0.527 0.597918
Kidhome                  1975.078    545.277   3.622 0.000299 ***
Teenhome                 5895.962    475.663  12.395  < 2e-16 ***
Recency                   -10.397      7.483  -1.389 0.164873
MntWines                   14.818      1.047  14.154  < 2e-16 ***
MntFruits                  13.510      7.567   1.786 0.074312 .
MntMeatProducts            22.138      1.661  13.328  < 2e-16 ***
MntFishProducts             4.165      5.750   0.724 0.468982
MntSweetProducts           27.121      7.291   3.720 0.000204 ***
MntGoldProds               -7.078      5.138  -1.378 0.168474
NumDealsPurchases        -515.367    145.162  -3.550 0.000393 ***
NumWebPurchases           999.450    109.659   9.114  < 2e-16 ***
NumCatalogPurchases      1018.332    130.341   7.813 8.61e-15 ***
NumStorePurchases         378.049    101.500   3.725 0.000201 ***
NumWebVisitsMonth       -2983.893    129.025 -23.127  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10150 on 2190 degrees of freedom
Multiple R-squared:  0.7794,    Adjusted R-squared:  0.7773
F-statistic: 368.5 on 21 and 2190 DF,  p-value: < 2.2e-16
```

- All numerical variables + Education and Marital_Status
- Four highest correlated variables are significant

Adj. $R^2$: **0.7773**

RSE: **10150**

p-value: **<0.05**

# Multiple Regression Models

| | Factors | $R^2$ | Adj. $R^2$ | RSE |
|---|---|---|---|---|
| Model 1 | MntWines<br>MntFruits<br>MntMeatProducts<br>MntFishProducts | **0.6338** | **0.6332** | **13030** |
| Model 2 | MntWines<br>MntMeatProducts<br>NumCatalogPurchases<br>NumWebVisitMonth | **0.7286** | **0.7281** | **11220** |
| Model 3 | MntWines<br>MntMeatProducts<br>NumCatalogPurchases<br>NumWebVisitMonth<br>Teenhome | **0.7527** | **0.7521** | **10710** |

# Income ~ MntWines * MntMeatProducts

```
Residuals:
   Min     1Q Median     3Q    Max
-32927  -7393   -147   6691 124638

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.304e+04  3.943e+02   83.81   <2e-16 ***
MntMeatProducts           7.907e+01  2.224e+00   35.55   <2e-16 ***
MntWines                  4.181e+01  1.211e+00   34.52   <2e-16 ***
MntMeatProducts:MntWines -7.410e-02  3.712e-03  -19.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12240 on 2208 degrees of freedom
Multiple R-squared:  0.6768,    Adjusted R-squared:  0.6764
F-statistic:  1541 on 3 and 2208 DF,  p-value: < 2.2e-16
```

- Interaction between meat and wine purchases is significant
- In context of our chosen model?

Adj. $R^2$: **0.6764**

RSE: **12240**

p-value: **<0.05**

# Model 3 + MntWines * MntMeatProducts

```
Residuals:
   Min     1Q Median     3Q    Max
-58444  -5918    -62   5491 108612

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.036e+04  8.065e+02  62.437   <2e-16 ***
MntWines                   3.505e+01  1.252e+00  27.995   <2e-16 ***
MntMeatProducts            5.622e+01  2.453e+00  22.920   <2e-16 ***
NumCatalogPurchases        2.224e+02  1.355e+02   1.642    0.101
NumWebVisitsMonth         -3.090e+03  1.089e+02 -28.374   <2e-16 ***
Teenhome                   5.231e+03  4.261e+02  12.277   <2e-16 ***
MntWines:MntMeatProducts -5.554e-02  3.453e-03 -16.083   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10140 on 2205 degrees of freedom
Multiple R-squared:  0.7787,    Adjusted R-squared:  0.7781
F-statistic:  1293 on 6 and 2205 DF,  p-value: < 2.2e-16
```

- NumCatalogPurchases no longer significant
  - Removing it

Adj. $R^2$: **0.7779**

RSE: **10140**

p-value: **<0.05**

# Model 4: Without NumCatalogPurchases

```
Residuals:
   Min     1Q Median     3Q    Max
-58631  -5789     62   5522 108287

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.066e+04  7.858e+02   64.46   <2e-16 ***
MntWines                   3.619e+01  1.044e+00   34.67   <2e-16 ***
MntMeatProducts            5.857e+01  1.997e+00   29.33   <2e-16 ***
NumWebVisitsMonth         -3.129e+03  1.062e+02  -29.46   <2e-16 ***
Teenhome                   5.204e+03  4.259e+02   12.22   <2e-16 ***
MntWines:MntMeatProducts -5.788e-02  3.145e-03  -18.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10140 on 2206 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7779
F-statistic:  1550 on 5 and 2206 DF,  p-value: < 2.2e-16
```

- All factors are significant
- Adj. $R^2$ matches that of our base model
- RSE is as small as our base model

Adj. $R^2$: **0.7781**

RSE: **10140**

p-value: **<0.05**

# Model Comparison

## Base Model

All Covariates

| | |
|---|---|
| Adj. $R^2$: | **0.7773** |
| RSE: | **10150** |
| p-value: | **<0.05** |

- First model used as baseline for comparison
- High adjusted $R^2$ value
- P-value indicates statistical significance

## Model 1

Spending Types

| | |
|---|---|
| Adj. $R^2$: | **0.6332** |
| RSE: | **13030** |
| p-value: | **<0.05** |

- Model of all spending options
- Lower adjusted $R^2$ than
- baseline
- Higher RSE value than baseline

## Model 2

Top 4 Highest Correlation Values

| | |
|---|---|
| Adj. $R^2$: | **0.7286** |
| RSE: | **11220** |
| p-value: | **<0.05** |

- Model of highest correlation
- Higher $R^2$ than M1, but less than baseline
- Lower RSE than M1, but higher than baseline

## Model 3

Correlation Values with TeenHome

| | |
|---|---|
| Adj. $R^2$: | **0.7521** |
| RSE: | **10710** |
| p-value: | **<0.05** |

- Model of highest correlation with teenhome
- Highest adjusted $R^2$ of the 3 models
- Lowest RSE of the 3
- Very close to baseline model

## Model 4

Model 3 + Interaction Term

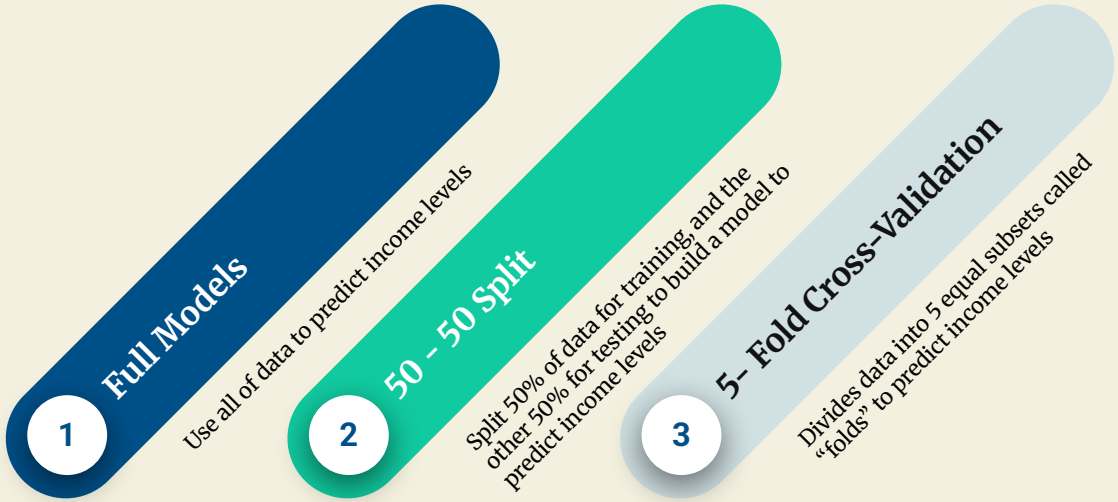
| | |
|---|---|
| Adj. $R^2$: | **0.7781** |
| RSE: | **10140** |
| p-value: | **<0.05** |

- Adds Interaction b/t MntMeatProducts and MntWines
- Lower RSE than baseline
- Higher Adj $R^2$ than baseline
- 6 terms vs 16 terms

# Predicting Income Levels

If Income < **$51, 301**
If Income > **$53,301**

**Low** Income -> Mapped to **0**
**High** Income -> Mapped to **1**




## Goal:

- Classify Income High or Low
- Same features used from regression models (MntWines,MntMeatProducts NumWebVists, Teenhome)
- Allows us to compare regression vs classification models
- Evaluating the models on accuracy, sensitivity, specificity and more.

# Models Using the Full Dataset



| Models | Accuracy | Sensitivity | Specificity | RunTime (s) |
|---|---|---|---|---|
| Logistic Regression | 0.913 | 0.895 | 0.931 | 0.040 |
| LDA | 0.897 | 0.849 | 0.946 | 0.048 |
| QDA | 0.896 | 0.832 | 0.959 | 0.069 |
| KNN | 0.994 | 0.994 | 0.999 | 1.2 |
| Naive Bayes | 0.894 | 0.833 | 0.954 | 0.051 |
| Decision Tree | 0.999 | 1.0 | 0.998 | 0.083 |
| Bagging | 0.993 | 0.996 | 0.990 | 0.072 |
| Random Forest | 0.999 | 0.998 | 1.0 | 0.617 |
| AdaBoost | 0.924 | 0.926 | 0.922 | 5.085 |
| XGBoost | 0.981 | 0.986 | 0.975 | 0.229 |

# Models Using 50% Validation



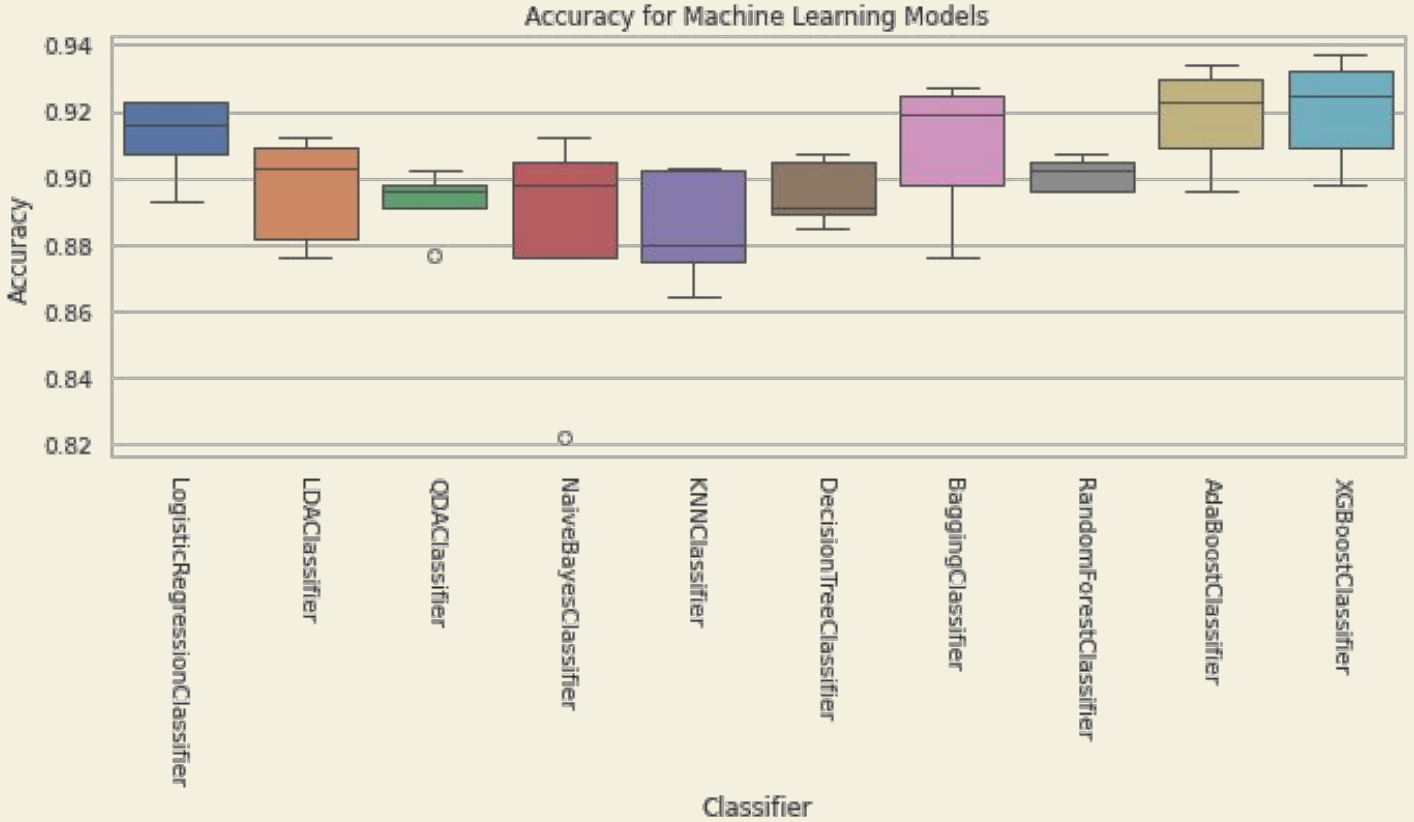
| Models | Accuracy | Sensitivity | Specificity | RunTime (s) |
|---|---|---|---|---|
| Logistic Regression | 0.907 | 0.864 | 0.952 | 0.037 |
| LDA | 0.900 | 0.850 | 0.949 | 0.031 |
| QDA | 0.892 | 0.852 | 0.931 | 0.041 |
| KNN | 0.863 | 0.869 | 0.858 | 0.894 |
| Naive Bayes | 0.892 | 0.856 | 0.928 | 0.036 |
| Decision Tree | 0.889 | 0.888 | 0.890 | 0.025 |
| Bagging | 0.891 | 0.909 | 0.876 | 0.113 |
| Random Forest | 0.896 | 0.895 | 0.896 | 0.398 |
| AdaBoost | 0.909 | 0.899 | 0.921 | 7.04 |
| XGBoost | 0.909 | 0.941 | 0.882 | 0.120 |

# Models Using Cross Validation

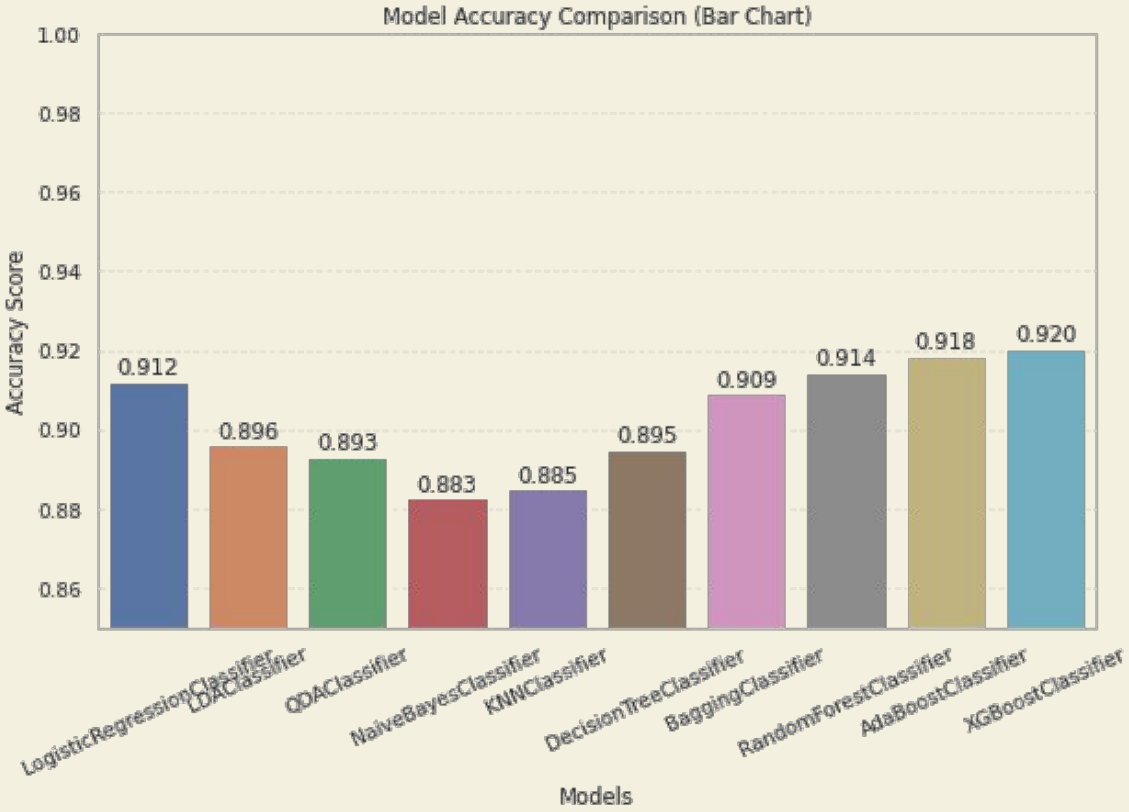


| Models | Accuracy | Sensitivity | Specificity | RunTime (s) |
|---|---|---|---|---|
| Logistic Regression | 0.914 | 0.899 | 0.929 | 27.194 |
| LDA | 0.896 | 0.848 | 0.945 | 2.413 |
| QDA | 0.893 | 0.829 | 0.957 | 2.498 |
| KNN (Optimal k = 7) | 0.885 | 0.897 | 0.872 | 0.206 |
| Naive Bayes | 0.894 | 0.835 | 0.954 | 3.760 |
| Decision Tree | 0.898 | 0.893 | 0.904 | 0.179 |
| Bagging | 0.908 | 0.919 | 0.898 | 0.323 |
| Random Forest (Optimal MF = 1) | 0.907 | 0.904 | 0.911 | 1.5 |
| AdaBoost | 0.918 | 0.920 | 0.917 | 16.82 |
| XGBoost | 0.920 | 0.914 | 0.926 | 0.929 |

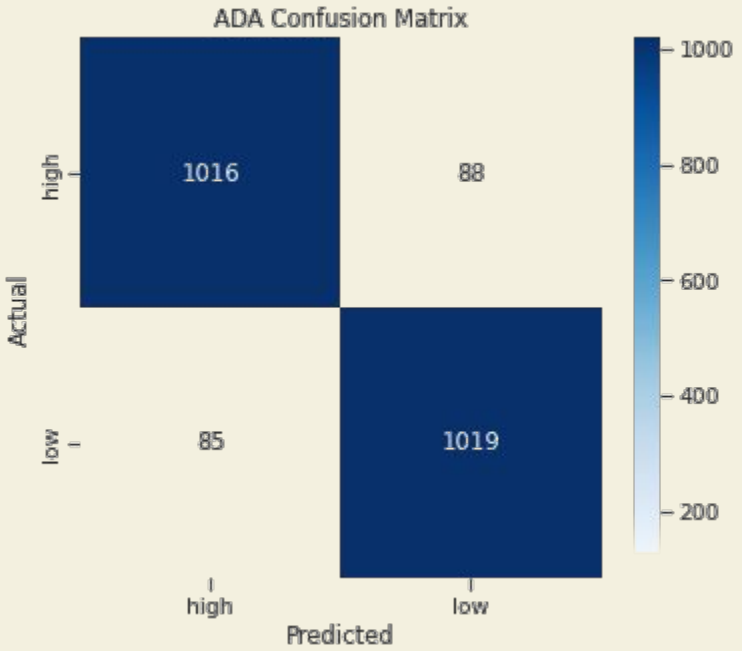
# Choosing the Best Models

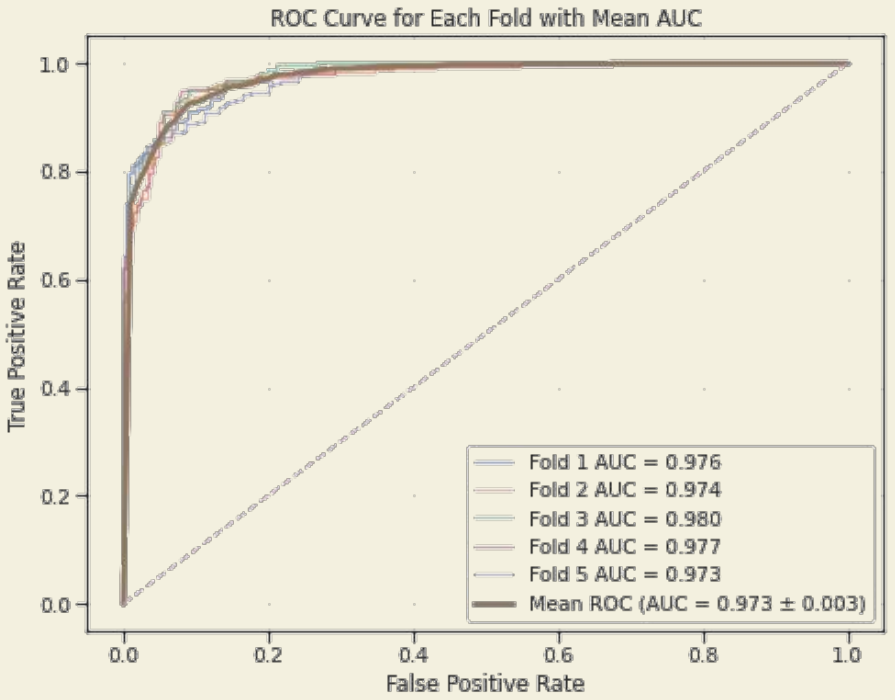


# Choosing the Best Models



# Best Model #1: ADA Boost

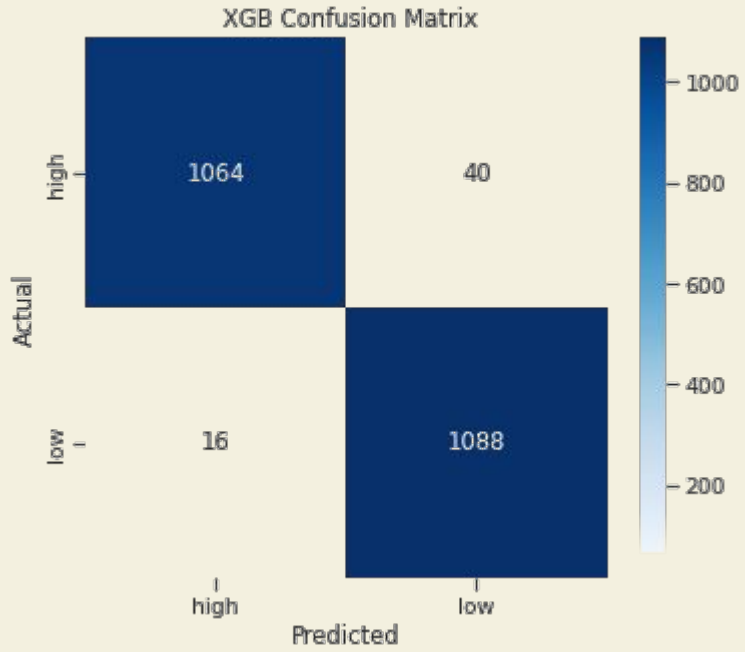


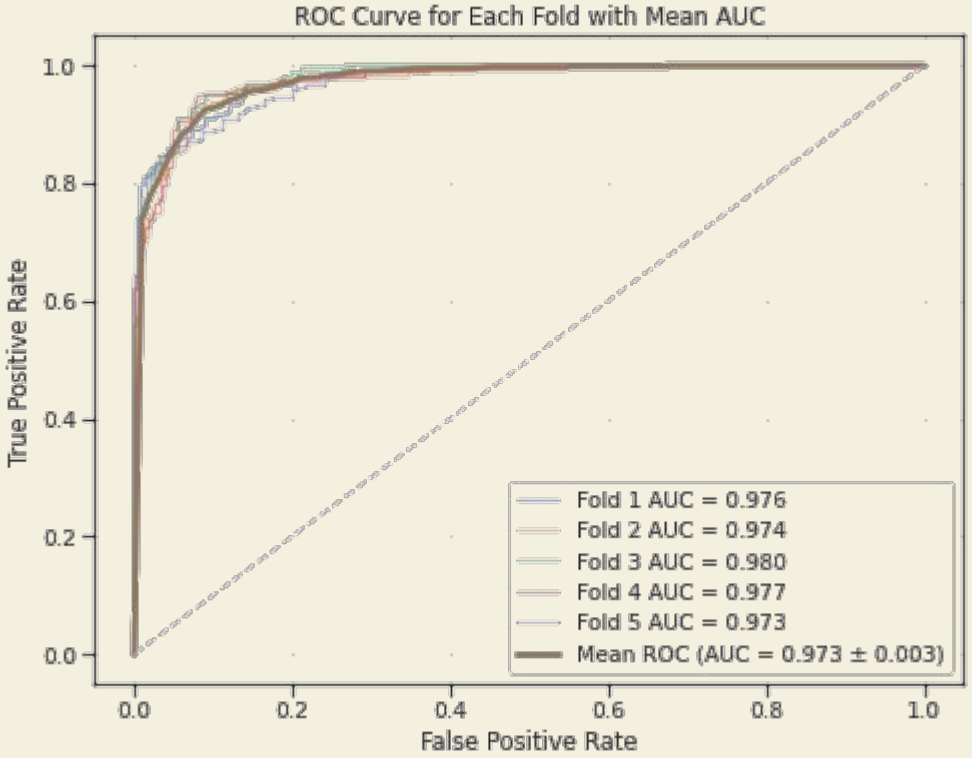



# Best Model #1: ADA Boost

| Model Type | Accuracy | Sensitivity | Specificity | SE | Precision | F1 - Score | AUC | Running Time (S) |
|---|---|---|---|---|---|---|---|---|
| Full Model | 0.924 | 0.927 | 0.922 | 0.006 | 0.926 | 0.924 | 0.981 | 1.686 |
| 50-50 Split | 0.909 | 0.899 | 0.921 | 0.014 | 0.886 | 0.903 | 0.975 | 2.141 |
| 5-Fold CV | 0.918 | 0.980 | 0.916 | 0.005 | 0.917 | 0.919 | 0.977 | 16.82 |

# Best Model #2: XGBoost



# Best Model #2: XGBoost

| Model Type | Accuracy | Sensitivity | Specificity | SE | Precision | F1 - Score | AUC | Running Time (S) |
|---|---|---|---|---|---|---|---|---|
| Full Model | 0.981 | 0.986 | 0.975 | 0.003 | 0.975 | 0.981 | 0.999 | 0.229 |
| 50-50 Split | 0.909 | 0.941 | 0.882 | 0.014 | 0.872 | 0.905 | 0.971 | 0.120 |
| 5-Fold CV | 0.920 | 0.926 | 0.915 | 0.006 | 0.916 | 0.921 | 0.975 | 0.929 |

# Comparing them

| Criteria | AdaBoost | XGBoost |
|---|---|---|
| Overall Strength | Fast, Simple, Stable | **Strong generalization & complex pattern detection** |
| Accuracy | High | **High** |
| Sensitivity | **Higher (Best At Detecting High Income)** | Moderate |
| Specificity | Moderate | **Higher (Fewer False Positives)** |
| CV Performance | Good | **Very Strong** |
| Runtime | Slowest | **Fast** |
| Model Complexity | Lower | **Higher** |

# Revisiting Goals

- With our data, we were looking to predict customer income based on different metrics (e.g. Customer Spending, Customer Response to Advertisement)
- We aimed to accomplish this using numerous regression models and classification methods.

# Regression & Classification

Performed EDA on dataset to find variables of highest correlation

Formed simple and multiple linear regression models to determine which variables are best to predict income

Found best model from regression analysis, then applied 10 different classifiers to the best model

Of those 10 classifiers we found that XGBoost and ADA Boost were best at predicting income