





# Predicting Customer Spending & Coupon Use from Dunnhumby Dataset



Capstone 3 Project  
Annie Erbsen  
December 2023



# Problem Statement

---

The Dunnhumby Complete Journey dataset contains household level transactions over two years from a group of 2,500 households that shop at a popular retailer.

*The primary objectives are to:*

- Understand how direct marketing and household demographics influence customer spending & coupon use
- Predict future customer spending and coupon use
- Make recommendations on how to develop more effective marketing campaigns based on my findings.

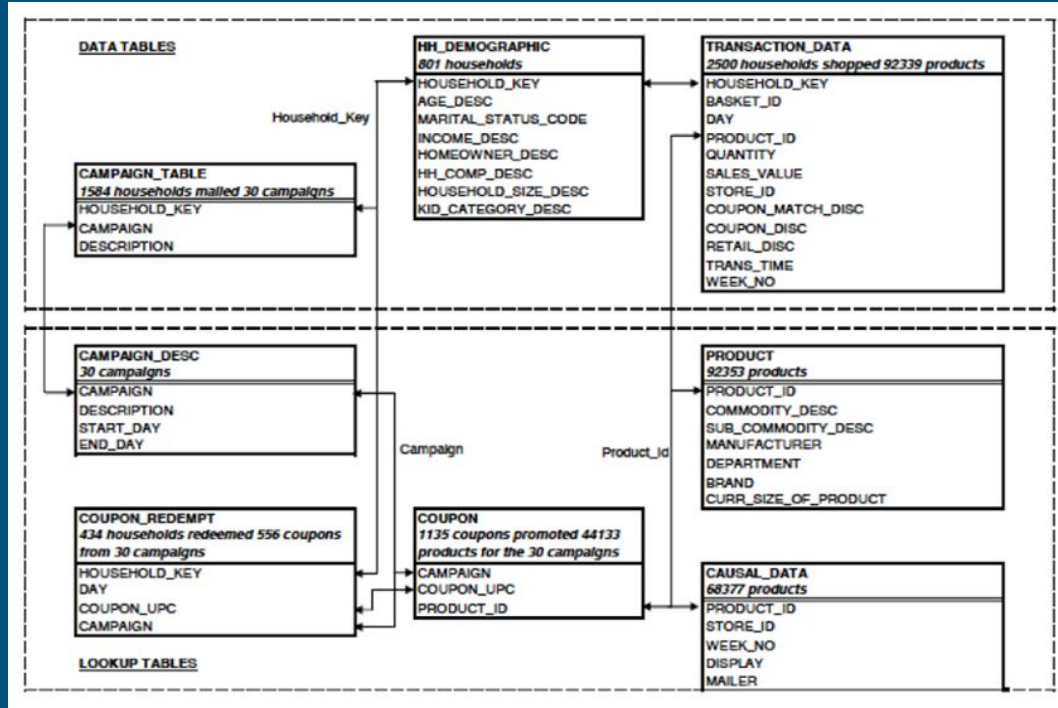
# The Data

Sourced from Kaggle: Dunnhumby - The Complete Journey

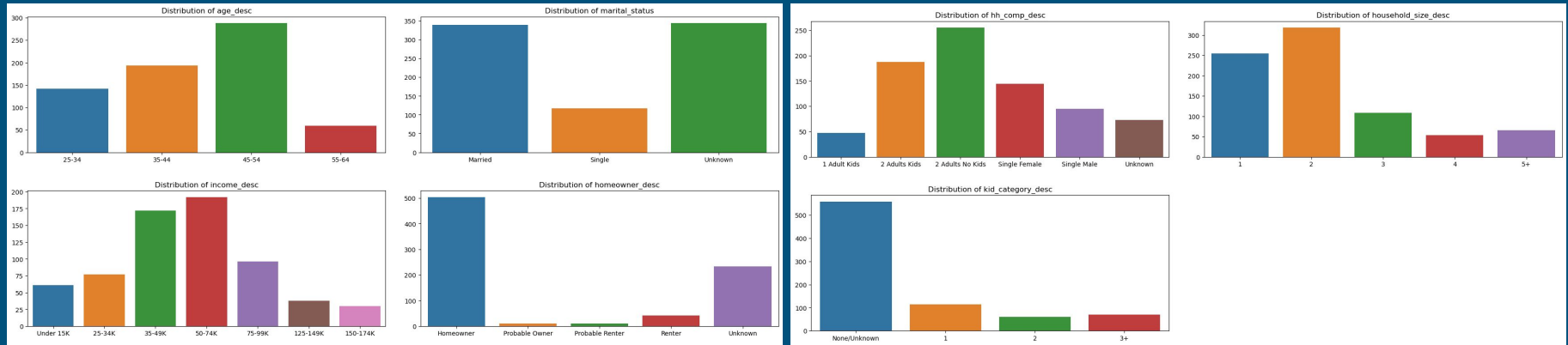
The dataset consists of 8 tables, detailed below:

	df_name	number_of_rows	number_of_columns	column_names
0	campaign_desc	30	4	DESCRIPTION, CAMPAIGN, START_DAY, END_DAY
1	campaign_table	7208	3	DESCRIPTION, household_key, CAMPAIGN
2	causal_data	36786524	5	PRODUCT_ID, STORE_ID, WEEK_NO, display, mailer
3	coupon	124548	3	COUPON_UPC, PRODUCT_ID, CAMPAIGN
4	coupon_redempt	2318	4	household_key, DAY, COUPON_UPC, CAMPAIGN
5	hh_demographic	801	8	AGE_DESC, MARITAL_STATUS_CODE, INCOME_DESC, HOMEOWNER_DESC, HH_COMP_DESC, HOUSEHOLD_SIZE_DESC, KID_CATEGORY_DESC, household_key
6	product	92353	7	PRODUCT_ID, MANUFACTURER, DEPARTMENT, BRAND, COMMODITY_DESC, SUB_COMMODITY_DESC, CURR_SIZE_OF_PRODUCT
7	transaction_data	2595732	12	household_key, BASKET_ID, DAY, PRODUCT_ID, QUANTITY, SALES_VALUE, STORE_ID, RETAIL_DISC, TRANS_TIME, WEEK_NO, COUPON_DISC, COUPON_MATCH_DISC

# Connecting the Tables Together

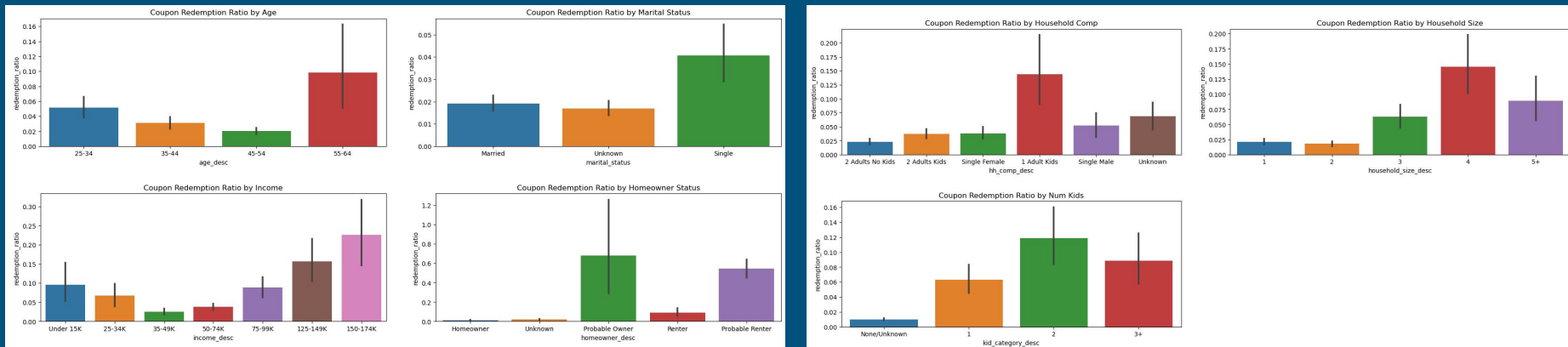


# Distribution of Customer Demographics



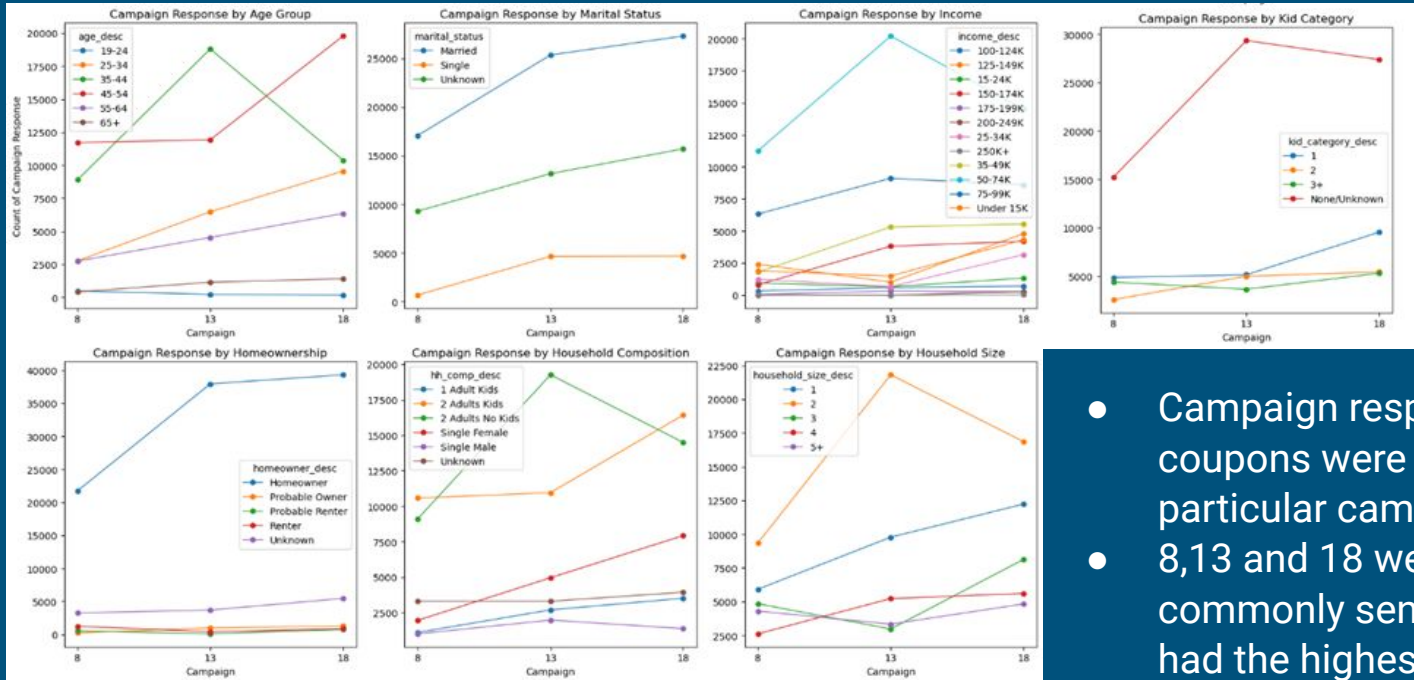
- Certain demographics were more widely included in this study.
- 'Unknown' marital status, homeowner status, household composition, and number of children make up a sizeable portion of the demographics.
- 801 of 2500 households have demographic data available. I included only data for the households with this data available.

# Coupon Redemption Rates



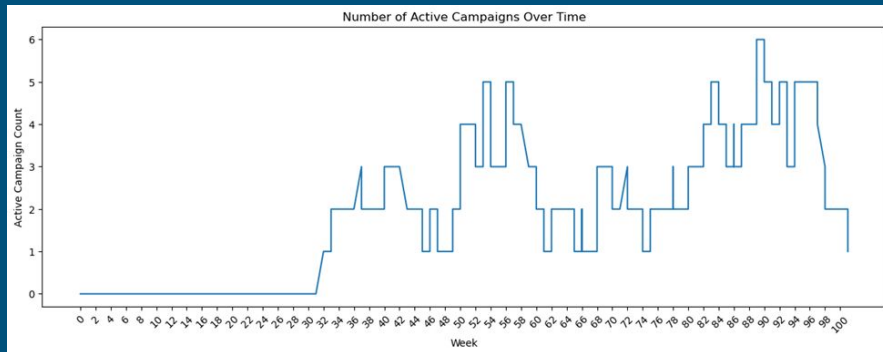
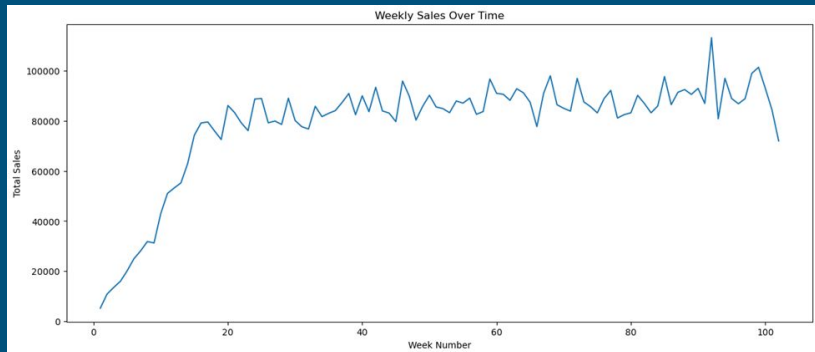
- The coupon redemption rates are, in some cases, inversely proportional to the population in the study.
- Coupon redemption rates are an important metric when considering which groups to send coupons to.
- Higher coupon redemption rates contribute to higher spending.

# Campaign Responses Across Demographics



- Campaign responses indicate that coupons were redeemed from a particular campaign.
- 8, 13 and 18 were the most commonly sent out campaign, and had the highest response rates.

# Weekly Sales & Active Campaigns Over Time

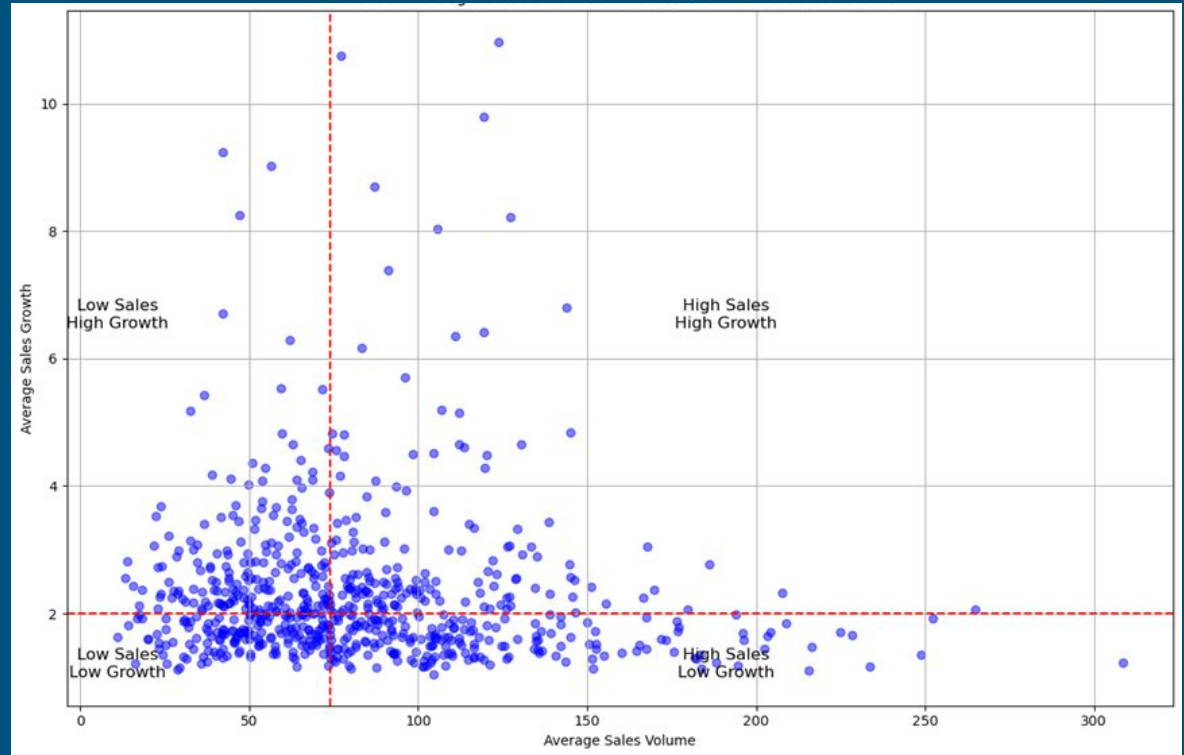


- Sales values were not fully collected until week 20.
- Campaigns were active after week 32.
- I kept the data when there were active campaigns.



# Household Segmentation Based on Sales Growth & Sales Volume

- High sales/high growth households should be top priority for marketing campaigns.
- Low sales/high growth and low high sales/low growth could also be target for campaigns.
- Low sales/low growth households could likely be excluded from campaigns as they are unlikely to respond.



# Modeling

---

I built 2 separate models:

- Weekly Sales Prediction Model (XGBoost Regression )
- Coupon Use Prediction Model (XGBoost Classifier)

# Sales Prediction Model

---

- Built an XGBoost Regression Model to predict weekly sales totals for households.
- The features included the customer demographics and coupon redemption rates.
- Due to outliers skewing the data, I built the model on a subset of the data with an interquartile range multiple of 0.25.
  - Model performance dropped when I ran the model on the entire dataset.
  - This model is not useful for predicting sales values, but it does tell us the directionality of the sales. SHAP values will provide useful information about how features contribute toward sales volume.

## **METRICS:**

### On 0.25 IQR subset of the data:

Mean Weekly Sales Value = \$58

Mean Absolute Error (MAE) = 28.96

Cross Validation MAE = 29.07

### On the full data range:

Mean Weekly Sales Value = \$80

Mean Absolute Error (MAE) = 45.93

Cross Validation MAE = 50.68

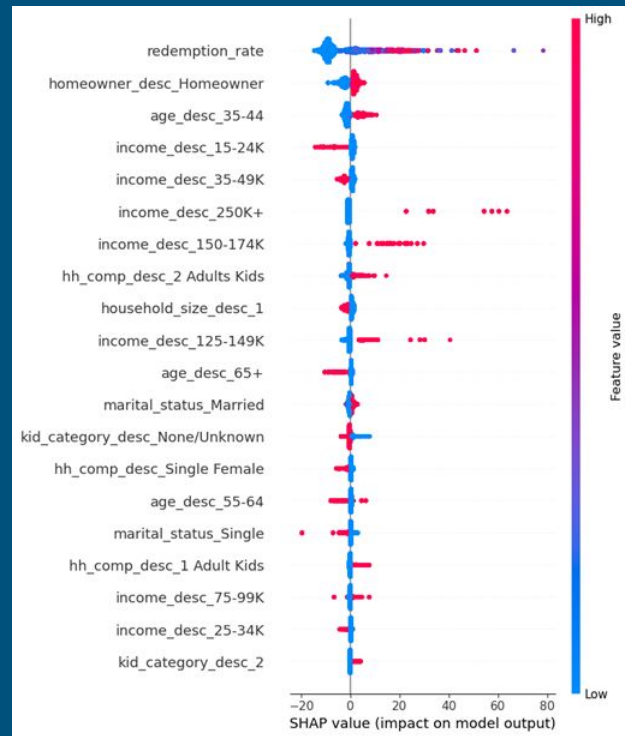
# SHAP Analysis for Sales Prediction Model

## Features that contribute towards higher sales:

- Higher redemption rates
- Homeowners earning over 24K a year
- 2 person households consisting of a married couple
- 35-44 year olds, especially if they earn 35-49K a year
- 75-99K a year
- 125-174K a year
- 250K+ a year
- 2 person households with kids who are not single
- Households composed of 1 adult and children, especially if they earn 35-49K a year
- 'Unknown' marital status with redemption rates over 1

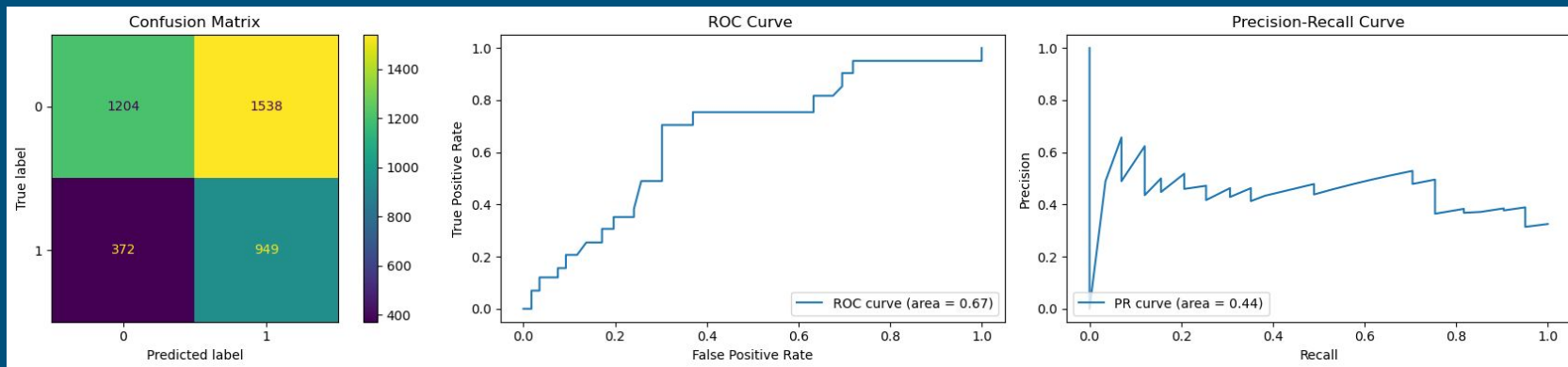
## Features that contribute towards lower sales:

- 15-24K income, especially if they are homeowners
- 35-49K income, especially if they are homeowners (except 35-44 year olds and single parents)
- 1 person households/single people, especially if they are homeowners and/or females
- Customers who are age 55+
- 'Unknown' homeowners



# Coupon Use Prediction Model

- Built XGBoost Classification Model to predict if a household will ever use a coupon or not based on demographics.
- To address imbalance classes, I used ADASYN oversampling.
- Removed features with negative permutation importance to improve model performance.
- **Metrics:** Positive Class Recall = 0.72, Cross Validation Recall = 0.75



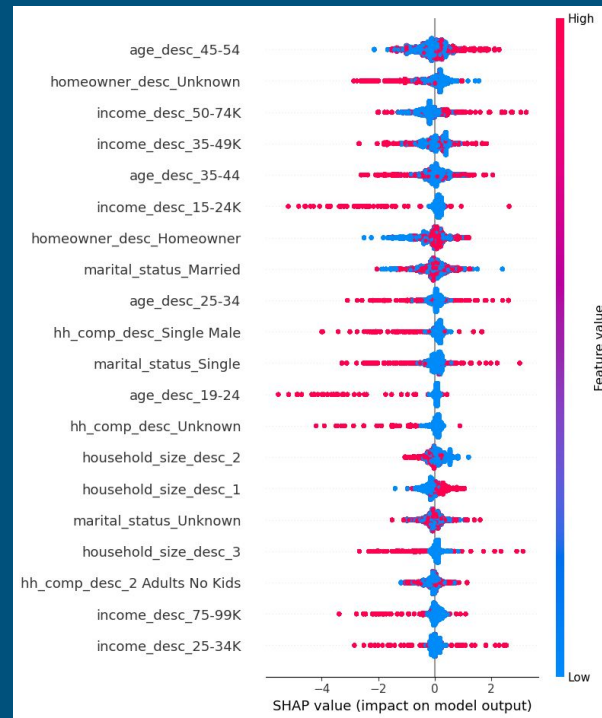
# SHAP Analysis for Coupon Use Prediction Model

## Features that contribute towards coupon use:

- Income 50-74K who are renters
- Homeowners who make under 15K
- Ages 25-34 with 2 kids
- Household size of 1 who make under 15K
- Household size of 3 who make 15-24K

## Features that contribute against coupon use:

- Ages 45-54 who make 35-49K a year
- Ages 35-44 with 2 kids
- Income 15-24K within the age bracket 45-54
- Married people with unknown homeowner status
- Single homeowners
- Income 15-34K who are age 35-44



# Conclusions

---

- Identified which demographics have higher overall response rates.
- Identified the sales growth and sales volume of each household, and segmented into quartiles.
- Sales Prediction model:
  - Not a useful model for predicting sales values.
  - Is a useful model for identifying which customer demographics contribute towards sales totals.
- Coupon Use Prediction Model
  - Correctly predicts which customers will use coupons 71% of the time, but with high false positive rate.
  - Is a useful model for identifying which customer demographics contribute toward coupon use.

# Future Work

---

- Do further analysis on seasonal trends, both at broad and product levels.
- Create time series forecasts for total sales and sales by product type.
- Investigate what is different about the 'total sales values' outliers, and to see if it makes sense to build separate models for customers who spend over 0.25 IQR.



# Recommendations

---

- Work with a marketing strategist to develop targeted marketing campaigns.
  - Identify which customers are likely to use coupons with the Coupon Use Prediction Model to develop a baseline list.
  - From there, can develop targeting marketing campaigns for different demographics identified in the Sales Prediction Model.
  - The combination of these models should result in a more cost-effective campaign with a higher response rate and increased sales.
- Send out marketing campaigns to customers in the study in the 'high growth/high sales' and 'high growth/low sales' lists as these are likely to be responsive to continued marketing efforts.