Capstone 3 Project Proposal:
Predicting and Analyzing Customer Spending Patterns using Dunnhumby Data

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 objective is to understand how direct marketing and household demographics influence customer spending, and to predict future customer spending to build more effective marketing campaigns.

Context:
Each household in this study has been sent multiple direct marketing campaigns over the course of 2 years. Each campaign consists of coupons that were sent to specific households that were valid during certain time periods. The dataset includes data on which coupons were redeemed by a household, for what product, the store location, and when. There are also data about what in-store displays were in specific stores over different time periods.. In addition to this, we have all of the purchase histories from each household from this retailer over the two year time span, as well as demographic data from 32% of the households. For the purchases, we have the day, the week and the time of day the purchase took place.

Criteria for Success:
- Accurate prediction of customer spending patterns within a reasonable margin of error (i.e. RMSE, MAE).
- Identifiable trends or impacts of direct marketing and demographics on spending.

Scope of Solution Space:
This project will involve preprocessing the data, EDA, feature engineering, model development, and time series analysis and forecasting. The project will be carried out in a database management system, most likely Snowflake. The deliverables will include a Tableau dashboard, a paper detailing my findings, a code repository and a presentation slide deck.

Constraints:
- The data is limited to 2500 households, and there is demographic information for only 801 of these households.
- The actual dates are not available; the days are numbered 1 through 711, and the weeks are numbered 1 through 102.
- We do not know the actual locations of the retailer, the overall income and cost of living demographics for the store locations, or details about the marketing campaigns aside from what customers used them for.

Stakeholders:

The stakeholders for this project are the retail marketing managers and the data science team.

Data Sources:

Data acquired from Dunnhumby's Complete Journey dataset on Kaggle: https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey. It includes 8 csv files that will be imported as tables into a database management system, and are summarized in the table 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

The data can all be easily joined together, as shown in the diagram below.