# Introduction

This code will accomplish the following task:

## Overall goal:

You're a marketing analyst and you've been told by the Chief Marketing Officer that recent marketing campaigns have not been as effective as they were expected to be. You need to analyze the data set to understand this problem and propose data-driven solutions.

## Section 01: Exploratory Data Analysis

* Are there any null values or outliers? How will you wrangle/handle them?
* Are there any variables that warrant transformations?
* Are there any useful variables that you can engineer with the given data?
* Do you notice any patterns or anomalies in the data? Can you plot them?

## Section 02: Statistical Analysis

Please run statistical tests in the form of regressions to answer these questions & propose data-driven action recommendations to your CMO. Make sure to interpret your results with non-statistical jargon so your CMO can understand your findings.

* What factors are significantly related to the number of store purchases?
* Does US fare significantly better than the Rest of the World in terms of total purchases?

Your supervisor insists that people who buy gold are more conservative. Therefore, people who spent an above average amount on gold in the last 2 years would have more in store purchases. Justify or refute this statement using an appropriate statistical test

Fish has Omega 3 fatty acids which are good for the brain. Accordingly, do "Married PhD candidates" have a significant relation with amount spent on fish? What other factors are significantly related to amount spent on fish? (Hint: use your knowledge of interaction variables/effects)

Is there a significant relationship between geographical regional and success of a campaign?

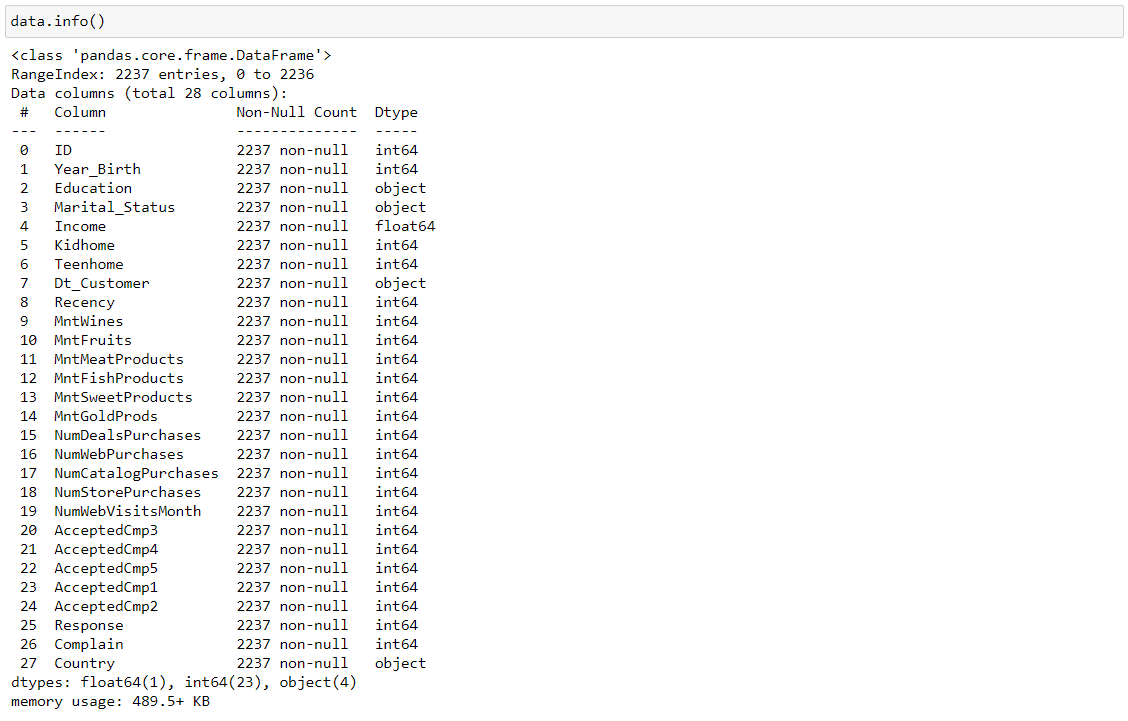
## Section 03: Data Visualization

Please plot and visualize the answers to the below questions.

* Which marketing campaign is most successful?
* What does the average customer look like for this company?
* Which products are performing best?
* Which channels are underperforming?

# Dataset

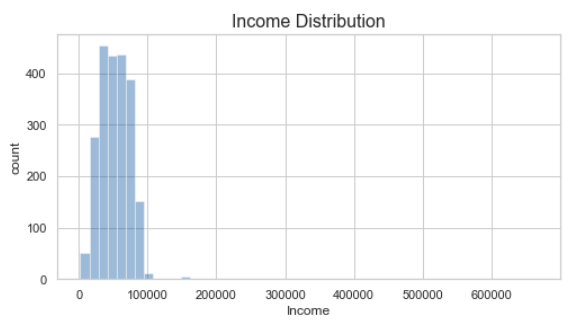
Before beginning the analysis, we will load and view the dataset, and perform some initial cleaning. Clean up column names, Transform selected columns to numeric format i.e., Income to float.



Identify features containing null values:



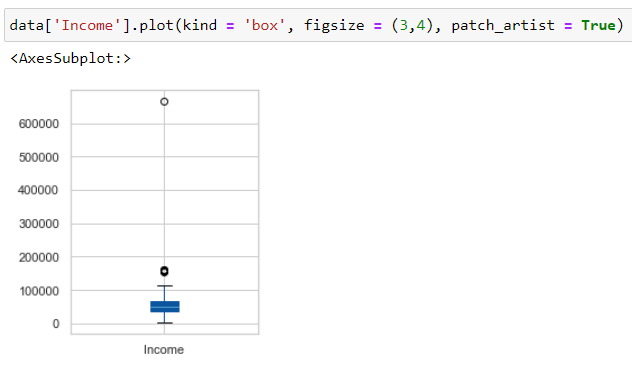
‘Income’ has 24 null values. Plotting the feature set to see the type of imputation to use.



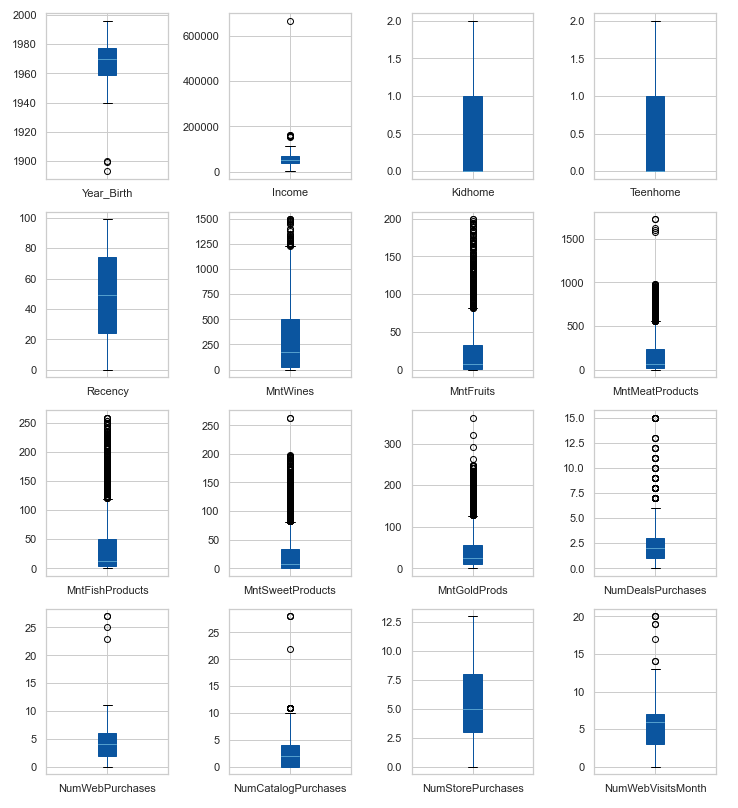
Findings:

* Most incomes are distributed between $0-\$100,000, with a few outliers
* Will impute null values with median value, to avoid effects of outliers on imputation value

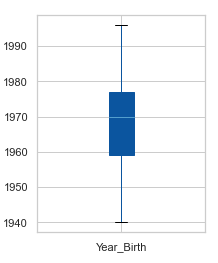
Plotting box plot for clarity on the outliers.



Impute null values in Income, using median value (to avoid skewing of the mean due to outliers). Boxplots provide a great insight for the outliers. Boxplots are made for all the features.

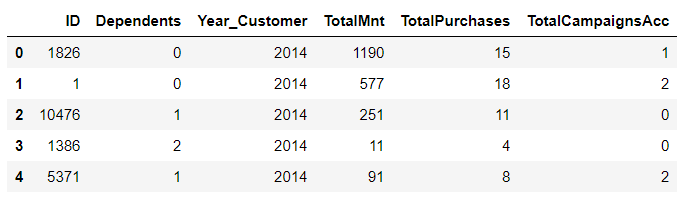


Findings: Multiple features contain outliers (see boxplots below), but the only that likely indicate data entry errors are Year\_Birth <= 1900. Removing instances where ‘Year\_Birth’ <= 1900.

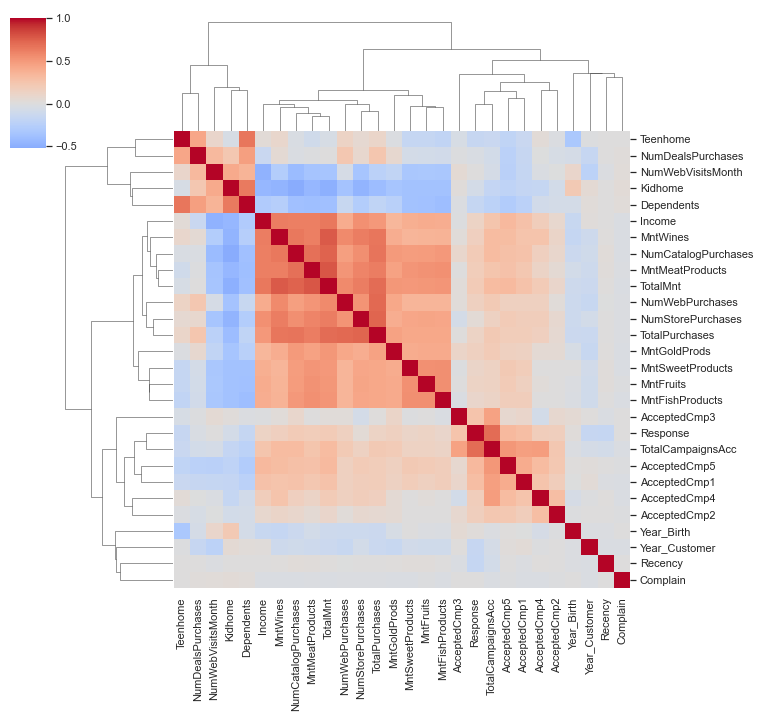


Transform Dt\_Customer to datetime since its data type is shown as an object. There is much useful data that can be salvaged from the dataset.

* The total number of dependents in the home ('Dependents') can be engineered from the sum of 'Kidhome' and 'Teenhome'
* The year of becoming a customer ('Year\_Customer') can be engineered from 'Dt\_Customer'
* The total amount spent ('TotalMnt') can be engineered from the sum of all features containing the keyword 'Mnt'
* The total purchases ('TotalPurchases') can be engineered from the sum of all features containing the keyword 'Purchases'
* The total number of campains accepted ('TotalCampaignsAcc') can be engineered from the sum of all features containing the keywords 'Cmp' and 'Response' (the latest campaign)



To identify patterns, we will first identify feature correlations. Positive correlations between features appear red, negative correlations appear blue, and no correlation appears grey in the clustered heatmap below.



From this heatmap we can observe the following clusters of correlated features:

The "High Income" cluster:

Amount spent ('TotalMnt' and other 'Mnt' features) and number of purchases ('TotalPurchases' and other 'Num... Purchases' features) are positively correlated with 'Income'

Purchasing in store, on the web, or via the catalog ('NumStorePurchases', 'NumWebPurchases', 'NumCatalogPurchases') is positively correlated with 'Income'

The "Have Kids & Teens" cluster:

Amount spent ('TotalMnt' and other 'Mnt' features) and number of purchases ('TotalPurchases' and other 'Num... Purchases' features) are negatively correlated with 'Dependents' (with a stronger effect from kids vs. teens)

Purchasing deals ('NumDealsPurchases') is positively correlated with 'Dependents' (kids and/or teens) and negatively correlated with 'Income'

The "Advertising Campaigns" cluster:

Acceptance of the advertising campaigns ('AcceptedCmp' and 'Response') are strongly positively correlated with each other

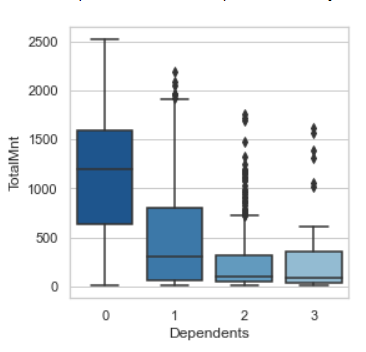
Weak positive correlation of the advertising campaigns is seen with the "High Income" cluster, and weak negative correlation is seen with the "Have Kids & Teens" cluster

Anomalies:

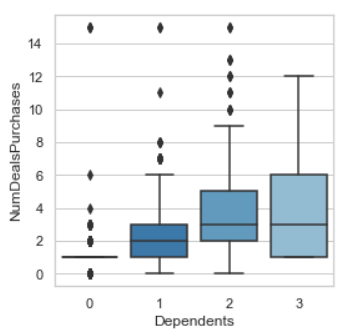
Surprisingly, the number of website visits in the last month ('NumWebVisitsMonth') does not correlate with an increased number of web purchases ('NumWebPurchases')

Instead, 'NumWebVisitsMonth' is positively correlated with the number of deals purchased ('NumDealsPurchases'), suggesting that suggesting that deals are an effective way of stimulating purchases on the website.

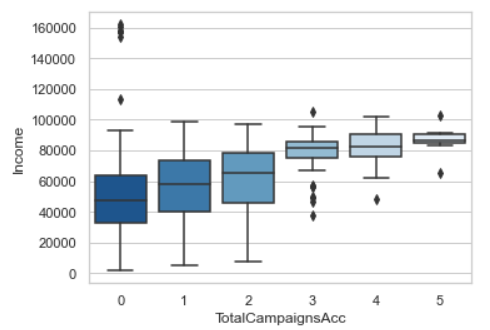
The following Plot illustrating negative effect of having dependents (kids & teens) on spending.

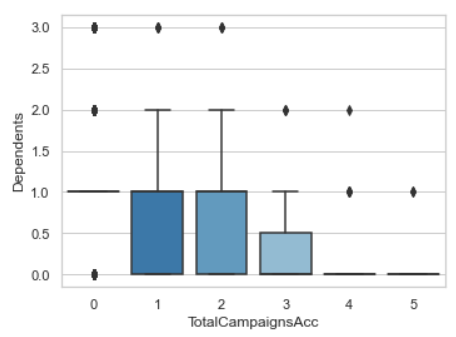


Plot illustrating positive effect of having dependents (kids & teens) on number of deals purchased.



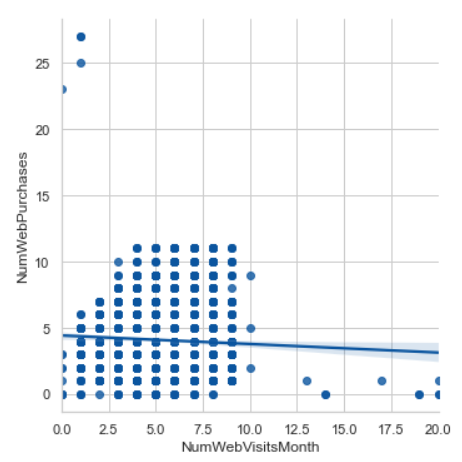
Plots illustrating the positive effect of income and negative effect of having kids & teens on advertising campaign acceptance.

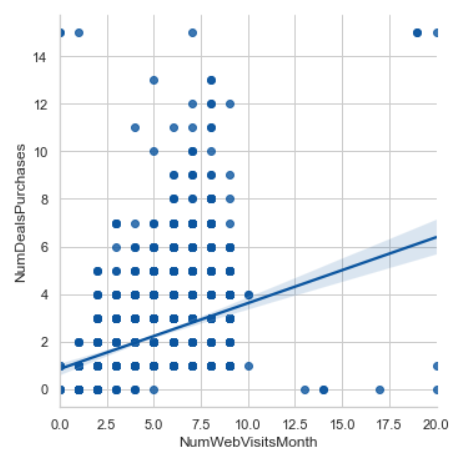




Investigate anomaly:

* Number of web visits in the last month is not positively correlated with number of web purchases
* Instead, it is positively correlated with the number of deals purchased, suggesting that deals are an effective way of stimulating purchases on the website





Running statistical tests in the form of regressions to answer questions & propose data-driven action recommendations.

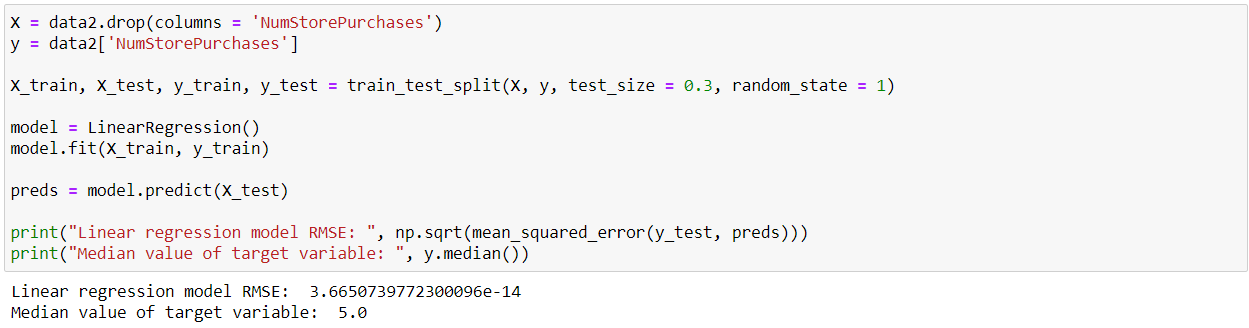
We will use use a linear regression model with NumStorePurchases as the target variable, and then use machine learning explainability techniques to get insights about which features predict the number of store purchases.



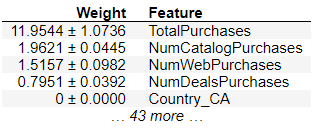
Drop uninformative features

* ID is unique to each customer
* Dt\_Customer will be dropped in favor of using engineered variable Year\_Customer

Perform one-hot encoding of categorical features. Fitting Linear regression on the dataset. The results are as follows.



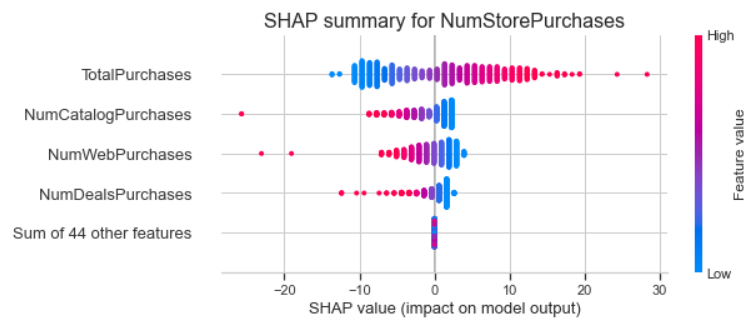
The RMSE is exceedingly small compared to the median value of the target variable, indicating good model predictions. Identify features that significantly affect the number of store purchases, using permutation importance:



* Significant features:
  + 'TotalPurchases', 'NumCatalogPurchases', 'NumWebPurchases', 'NumDealsPurchases'
  + All other features are not significant

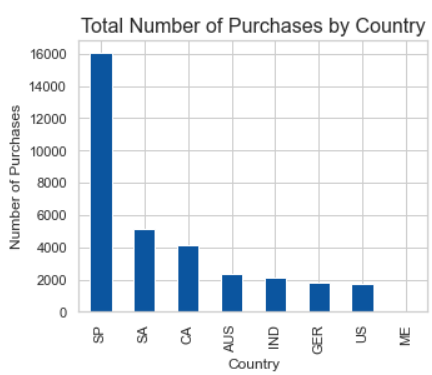
Explore the directionality of these effects, using SHAP values:

* Findings:
  + The number of store purchases increases with higher number of total purchases ('TotalPurchases')
  + The number of store purchases decreases with higher number of catalog, web, or deals purchases ('NumCatalogPurchases', 'NumWebPurchases', 'NumDealsPurchases')
* Interpretation:
  + Customers who shop the most in stores are those who shop less via the catalog, website, or special deals



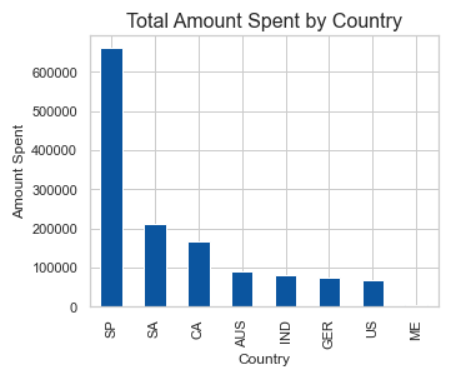
Does US fare significantly better than the Rest of the World in terms of total purchases?

* Plot total number of purchases by country:
* Findings:
  + Spain (SP) has the highest number of purchases
  + US is second to last, therefore the US does not fare better than the rest of the world in terms of the total number of purchases



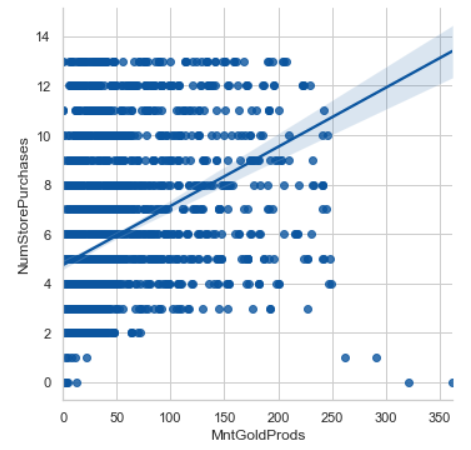
Plot total amount spent by country:

* Findings:
  + Spain (SP) has the highest total amount spent on purchases
  + US is second to last, therefore the US does not fare better than the rest of the world in terms of the total amount spent on purchases

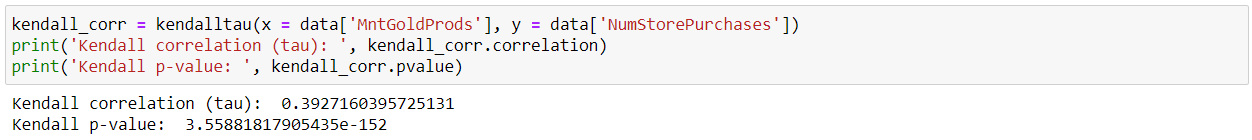


People who buy gold are more conservative. Therefore, people who spent an above average amount on gold in the last 2 years would have more in store purchases. Justify or refute this statement using an appropriate statistical test.

Plot relationship between amount spent on gold in the last 2 years (MntGoldProds) and number of in store purchases (NumStorePurchases).



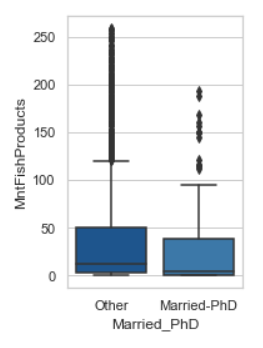
Perform Kendall correlation analysis (non-parametric test since MntGoldProducts is not normally distributed and contains outliers). Findings: There is significant positive correlation between MntGoldProds and NumStorePurchases.



Fish has Omega 3 fatty acids which are good for the brain. Accordingly, do "Married PhD candidates" have a significant relation with amount spent on fish?

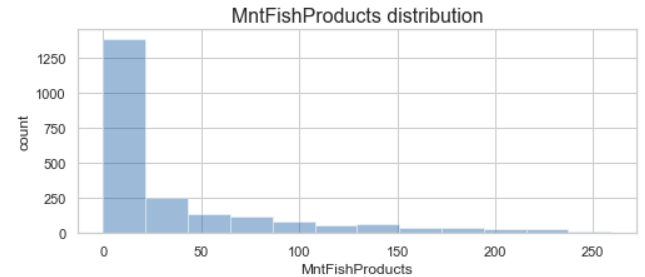
We will compare MntFishProducts between Married PhD candidates and all other customers:

Findings: Married PhD candidates spend significantly less on fish products compared to other customers.



What other factors are significantly related to amount spent on fish?¶

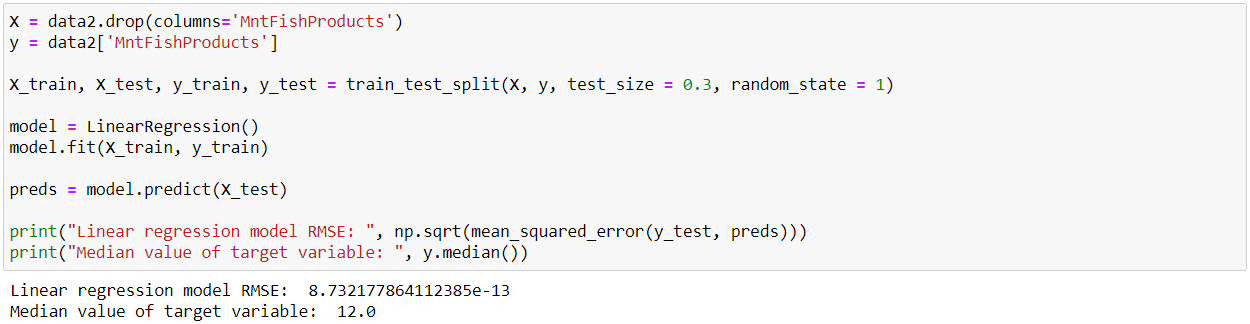
Like with the analysis of NumStorePurchases above, we will use use a linear regression model with MntFishProducts as the target variable, and then use machine learning explainability techniques to get insights about which features predict the amount spent on fish.



Fit linear regression model to training data (70% of dataset)

Evaluate predictions on test data (30% of dataset) using RMSE:

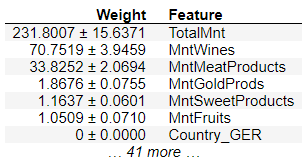
Findings: The RMSE is exceedingly small compared to the median value of the target variable, indicating good model predictions.



Identify features that significantly affect the amount spent on fish, using permutation importance:

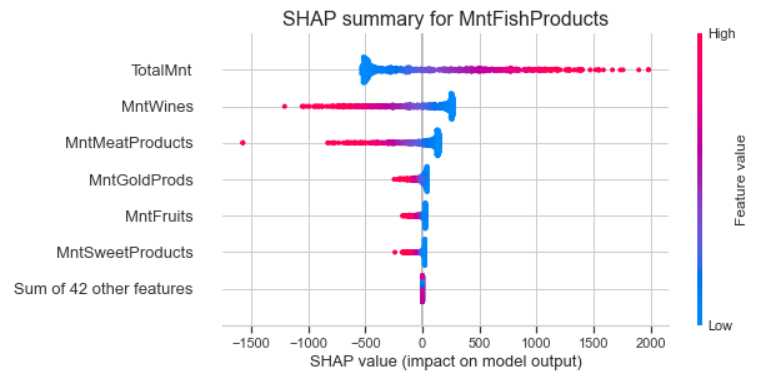
Significant features:

* 'TotalMnt', 'MntWines', 'MntMeatProducts', 'MntGoldProds', 'MntSweetProducts', 'MntFruits'
* All other features are not significant



Explore the directionality of these effects, using SHAP values:

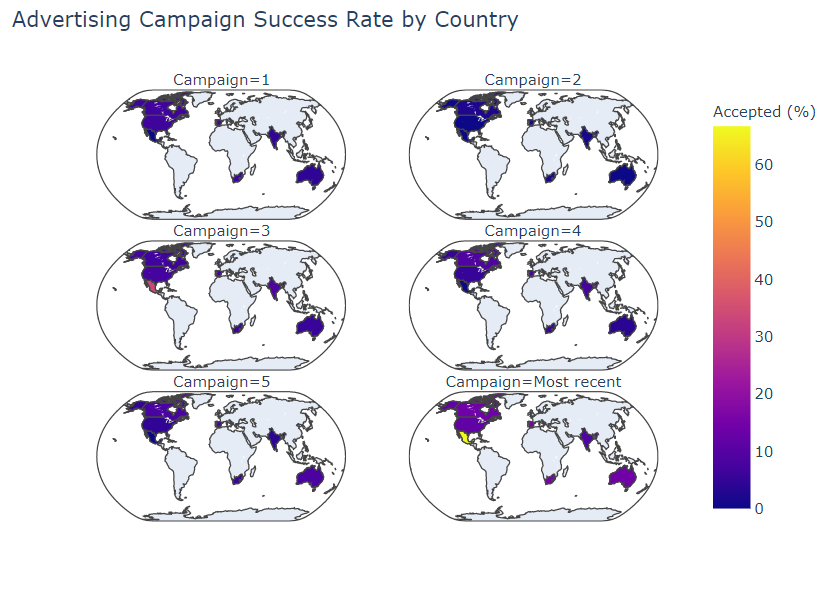
* Findings:
  + The amount spent on fish increases with higher total amount spent ('TotalMnt')
  + The amount spent on fish decreases with higher amounts spent on wine, meat, gold, fruit, or sweets ('MntWines', 'MntMeatProducts', 'MntGoldProds', 'MntSweetProducts', 'MntFruits')
* Interpretation:
  + Customers who spend the most on fish are those who spend less on other products (wine, meat, gold, fruit, and sweets)



Is there a significant relationship between geographical regional and success of a campaign?

Plot success of campaigns by region:

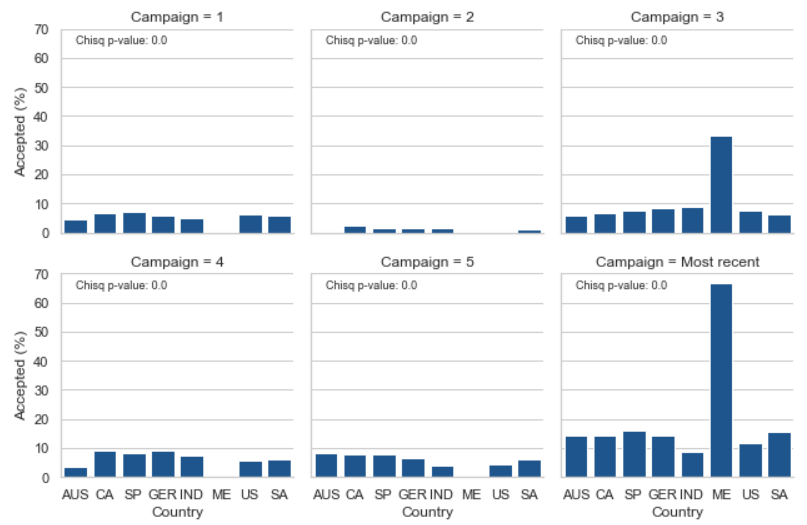
* Findings:
  + The campaign acceptance rates are low overall
  + The campaign with the highest overall acceptance rate is the most recent campaign (column name: Response)
  + The country with the highest acceptance rate in any campaign is Mexico
* Is the effect of region on campaign success statistically significant? See below.



Statistical summary of regional effects on campaign success:

Methodology: Performed logistic regression for Campaign Accepted by Country, reporting Chisq p-value for overall model.

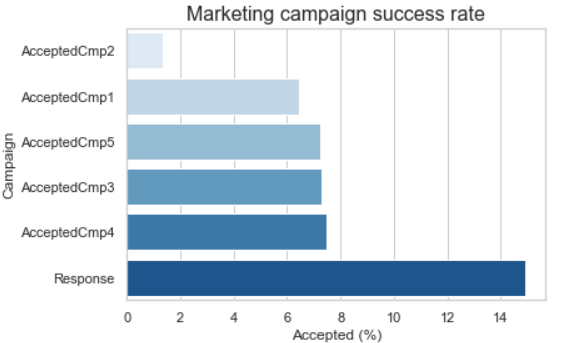
Findings: The regional differences in advertising campaign success are statistically significant.



Which marketing campaign is most successful?

Plot marketing campaign overall acceptance rates:

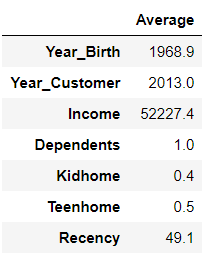
Findings: The most successful campaign is the most recent (column name: Response)



What does the average customer look like for this company?

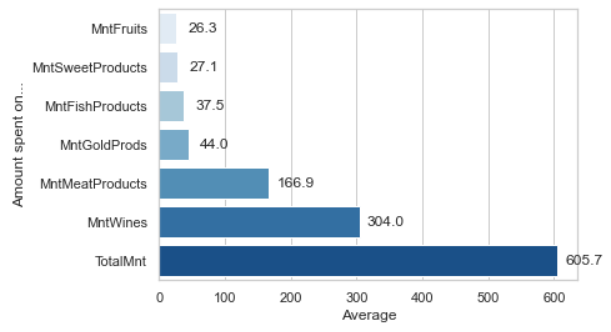
Basic demographics: The average customer is:

* Born in 1969
* Became a customer in 2013
* Has an income of roughly $52,000 per year
* Has 1 dependent (roughly equally split between kids or teens)
* Made a purchase from our company in the last 49 days



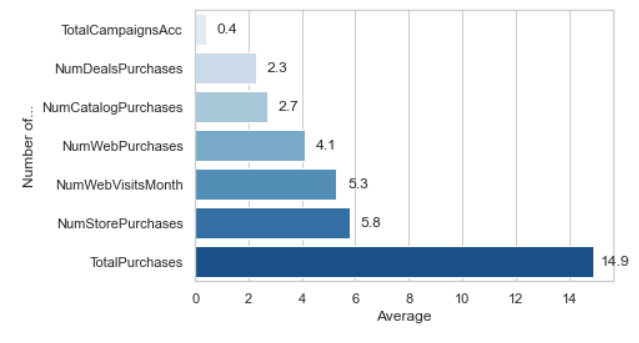
Which products are performing best?

* The average customer spent...
  + $25-50 on Fruits, Sweets, Fish, or Gold products
  + Over $160 on Meat products
  + Over $300 on Wines
  + Over $600 total
* Products performing best:
  + Wines
  + Followed by meats



Which channels are underperforming?

* Channels: The average customer:
  + Accepted less than 1 advertising campaign
  + Made 2 deals purchases, 2 catalog purchases, 4 web purchases, and 5 store purchases
  + Averaged 14 total purchases
  + Visited the website 5 times
* Underperforming channels:
  + Advertising campaigns
  + Followed by deals, and catalog



Summary of actionable findings to improve advertising campaign success:

The most successful advertising campaign was the most recent campaign (column name: Response), and was particularly successful in Mexico (>60% acceptance rate!)

Suggested action: Conduct future advertising campaigns using the same model recently implemented in Mexico.

Advertising campaign acceptance is positively correlated with income and negatively correlated with having kids/teens

Suggested action: Create two streams of targeted advertising campaigns, one aimed at high-income individuals without kids/teens and another aimed at lower-income individuals with kids/teens

The most successful products are wines and meats (i.e. the average customer spent the most on these items)

Suggested action: Focus advertising campaigns on boosting sales of the less popular items

The underperforming channels are deals and catalog purchases (i.e. the average customer made the fewest purchases via these channels)

The best performing channels are web and store purchases (i.e. the average customer made the most purchases via these channels)

Suggested action: Focus advertising campaigns on the more successful channels, to reach more customers