BUS AN 512 Data Management and SQL
Group Project

iFood CRM Case Study

Presenters:

JANN Ang

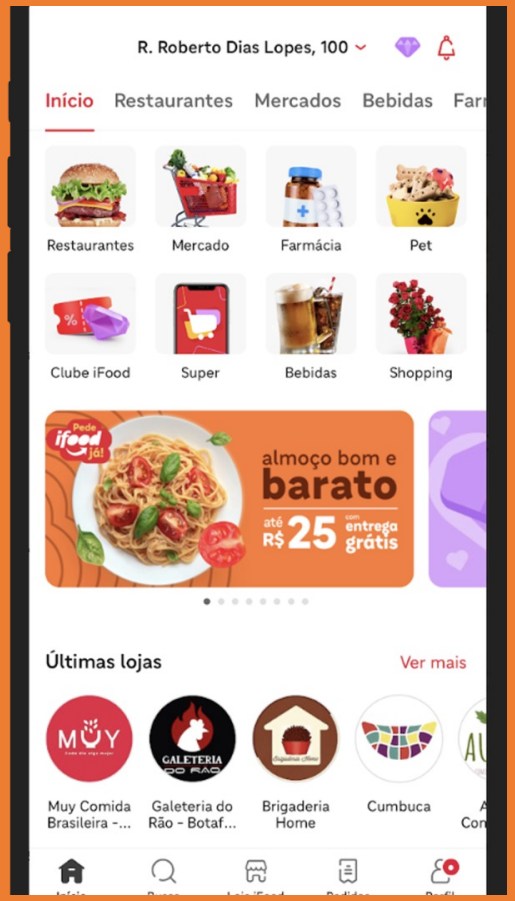
MAGGIE Ding

AISHWARYA Panse

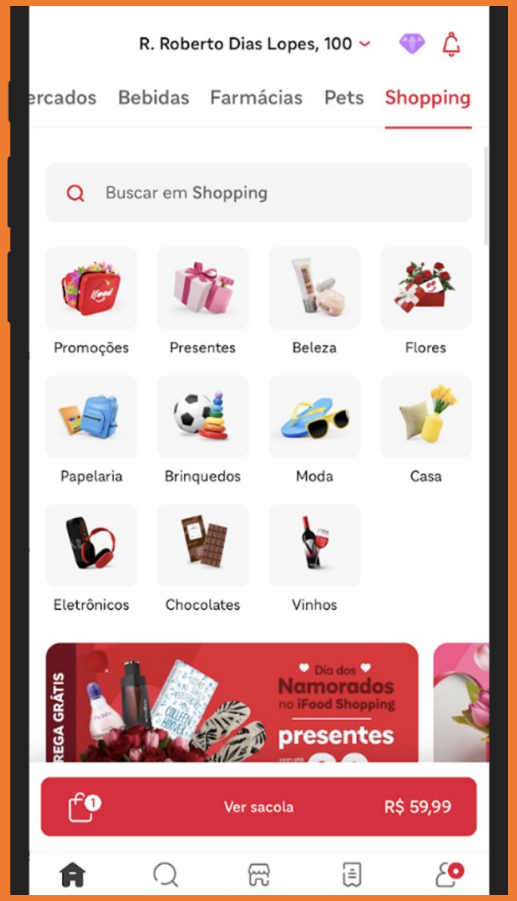
QUEENIE Chao

FARIN Fukunaga

iFood’s home screen



iFood has a variety of categories for different products



“Who wants a coupon?” is the title of the app’s promotions page

Description

Data's origin

Case study competition "iFood Brain team data challenge" provided by the iFood

With this sample dataset, we will analyze meta information on the customer and on iFood campaign interactions with their customers

Our challenge is understanding the data, finding business opportunities & insights, and analyze different segment of customer's spending

Key Objective

Explore the data – Provide insights, and define cause and effect. Provide a better understanding of the characteristic features of respondents

Describe customer segmentation based on customers' behaviors

Create 2 predictive models which allow the company to maximize the profit of their next marketing campaign

Dataset Overview

2240 rows x 20 columns

Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases
1957	Graduation	Single	58138	0	0	4/9/12	58	635	88	546	172	88	88	3	8	10
1954	Graduation	Single	46344	1	1	8/3/14	38	11	1	6	2	1	6	2	1	1
1965	Graduation	Together	71613	0	0	21/8/13	26	426	49	127	111	21	42	1	8	2
1984	Graduation	Together	26646	1	0	10/2/14	26	11	4	20	10	3	5	2	2	0
1981	PhD	Married	58293	1	0	19/1/14	94	173	43	118	46	27	15	5	5	3
1967	Master	Together	62513	0	1	9/9/13	16	520	42	98	0	42	14	2	6	4
1971	Graduation	Divorced	55635	0	1	13/11/12	34	235	65	164	50	49	27	4	7	3
1985	PhD	Married	33454	1	0	8/5/13	32	76	10	56	3	1	23	2	4	0
1974	PhD	Together	30351	1	0	6/6/13	19	14	0	24	3	3	2	1	3	0
1950	PhD	Together	5648	1	1	13/3/14	68	28	0	6	1	1	13	1	1	0
1983	Graduation	Married	NULL	1	0	15/11/13	11	5	5	6	0	2	1	1	1	0
1976	Basic	Married	7500	0	0	13/11/12	59	6	16	11	11	1	16	1	2	0
1959	Graduation	Divorced	63033	0	0	15/11/13	82	194	61	480	225	112	30	1	3	4
1952	Master	Divorced	59354	1	1	15/11/13	53	233	2	53	3	5	14	3	6	1
1987	Graduation	Married	17323	0	0	10/10/12	38	3	14	17	6	1	5	1	1	0
1946	PhD	Single	82800	0	0	24/11/12	23	1006	22	115	59	68	45	1	7	6
1980	Graduation	Married	41850	1	1	24/12/12	51	53	5	19	2	13	4	3	3	0
1946	Graduation	Together	37760	0	0	31/8/12	20	84	5	38	150	12	28	2	4	1
1949	Master	Married	76995	0	1	28/3/13	91	1012	80	498	0	16	176	2	11	4
1985	2n Cycle	Single	33812	1	0	3/11/12	86	4	17	19	30	24	39	2	2	1
1982	Graduation	Married	37040	0	0	8/8/12	41	86	2	73	69	38	48	1	4	2
1979	Graduation	Married	2447	1	0	6/1/13	42	1	1	1725	1	1	1	15	0	28
1949	PhD	Married	58607	0	1	23/12/12	63	867	0	86	0	0	19	3	2	3
1954	PhD	Married	65324	0	1	11/1/14	0	384	0	102	21	32	5	3	6	2
1951	Graduation	Together	40689	0	1	18/3/13	69	270	3	27	39	6	99	7	7	1
1969	Graduation	Single	18589	0	0	2/1/13	89	6	4	25	15	12	13	2	2	1
1976	Graduation	Married	53359	1	1	27/5/13	4	173	4	30	3	6	41	4	5	1
1986	Graduation	Single	NULL	1	0	20/2/13	19	5	1	3	3	263	362	0	27	0
1989	Graduation	Married	38360	1	0	31/5/13	26	36	2	42	20	21	10	2	2	1
1965	PhD	Married	84618	0	0	22/11/13	96	684	100	801	21	66	0	1	6	9
1989	Master	Divorced	10979	0	0	22/5/14	34	8	4	10	2	2	4	2	3	0
1963	Master	Together	38620	0	0	11/5/13	56	112	17	44	34	22	89	1	2	5
1970	Graduation	Together	40548	0	1	10/10/12	31	110	0	5	2	0	3	2	2	1
1952	PhD	Divorced	46610	0	2	29/10/12	8	96	12	96	33	22	43	6	4	1
1946	Master	Married	68657	0	0	20/2/13	4	482	34	471	119	68	22	1	3	5
1951	Master	Single	49389	1	1	29/8/13	55	40	0	19	2	1	3	1	2	0

Data Dictionary

Products from 5 major categories

Further divided into:

3 sales channels:

Feature	Description
ID	Unique identifier for each customer
Year_Birth	Customer's year of birth
DtCustomer	Date of customer's enrollment with the company
Education	Customer's level of education
Marital	Customer's marital status
Kidhome	Number of small children in customer's household
Teenhome	Number of teenagers in customer's household
Income	Customer's yearly household income
MntFishProducts	Amount spent on fish products in the last 2 years
MntMeatProducts	Amount spent on meat products in the last 2 years
MntFruits	Amount spent on fruit products in the last 2 years
MntSweetProducts	Amount spent on sweet products in the last 2 years
MntWines	Amount spent on wine products in the last 2 years
MntGoldProds	Amount spent on gold products in the last 2 year
NumDealsPurchases	Number of purchases made with discount
NumCatalogPurchases	Number of purchases made using catalogue
NumStorePurchases	Number of purchases made directly in stores
NumWebPurchases	Number of purchases made through company's website
NumWebVisitsMonth	Number of visits to company's website in the last month
Recency	Number of days since the last purchase

Data Cleaning

```
(select *,
case
when Income = '' then null
else Income
end as Income_clean
from ifood_data) as sub1)
```

1. Missing values: Some columns may have missing values represented as NULL using CASE statement
2. Dummy variables on the amount spent (wine, meat, fruit, etc): <= $10 : 0 and >$10 : 1

Well-documented 5

Well-maintained 5

Clean data 5

Challenges

- We can only select a few datasets that are significant/interesting enough to analyze it

Useful Data to have...

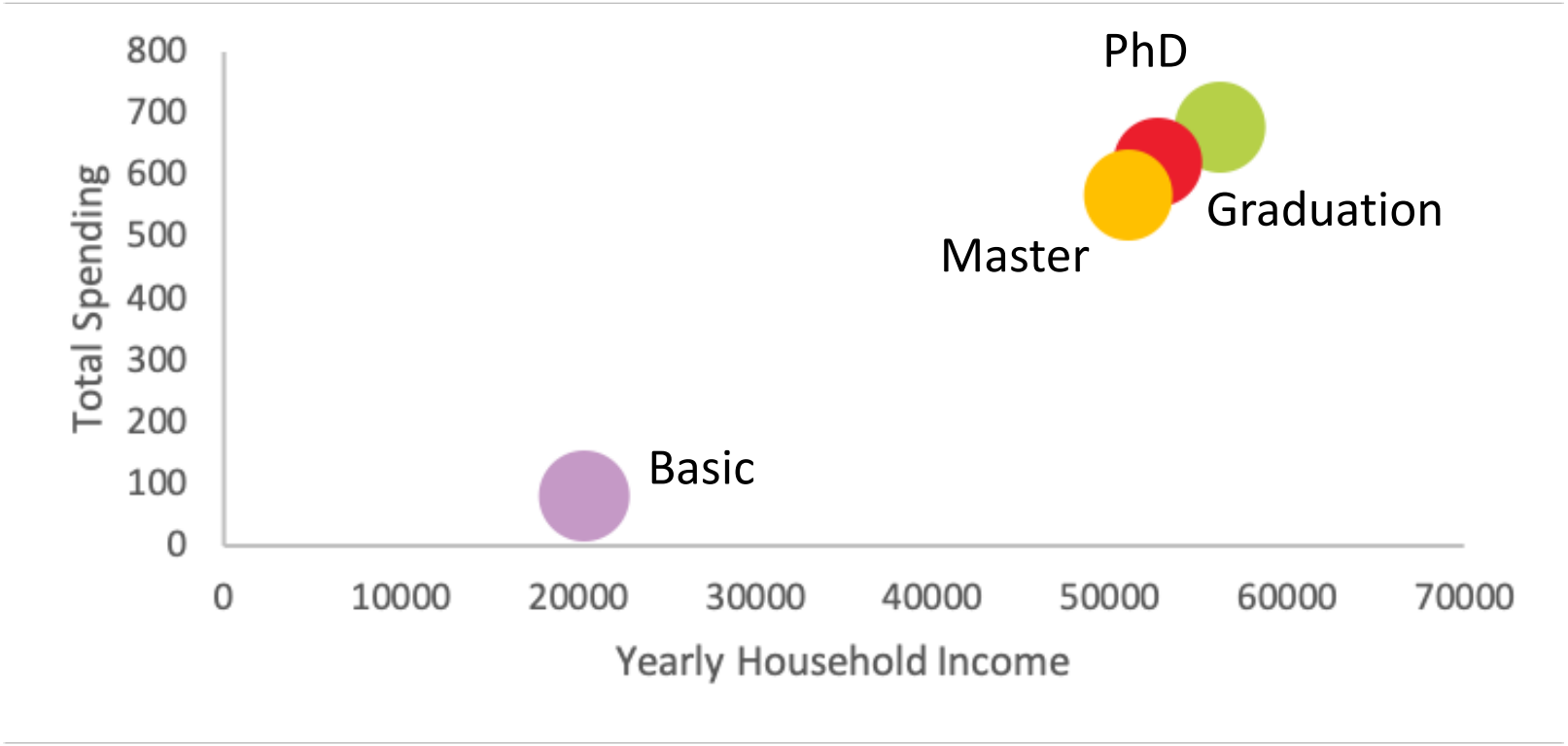
The region of Brazil these people came from

What exact products customer brought

What promotions customers responded to

Does a person's education level affect buying decisions?

How Education Level Can Influence Grocery Spending?



Higher Education Level

↓

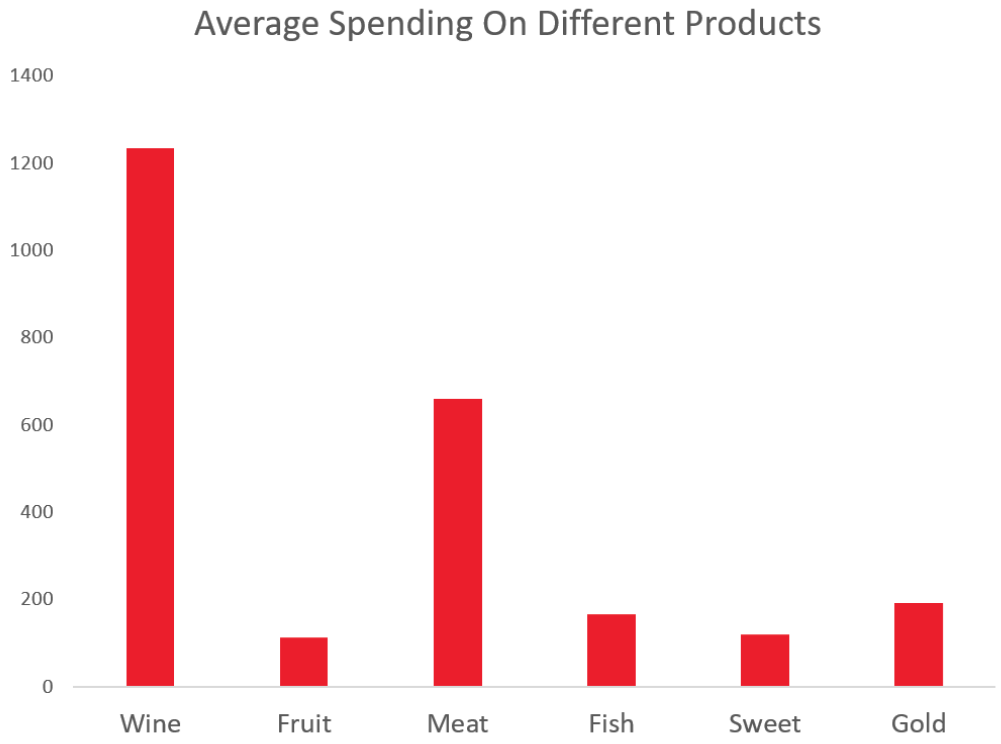
Higher Income

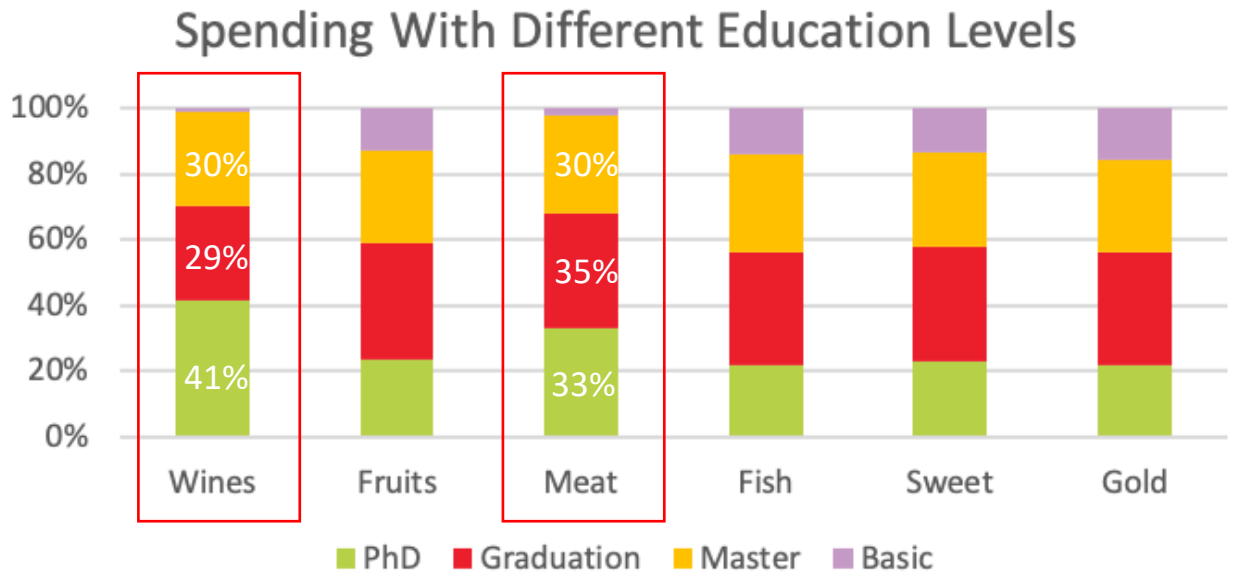
↓

More Spending

How Education Level Can Influence Grocery Spending?

Customers with higher education levels are more willing to spend more on higher-priced groceries.





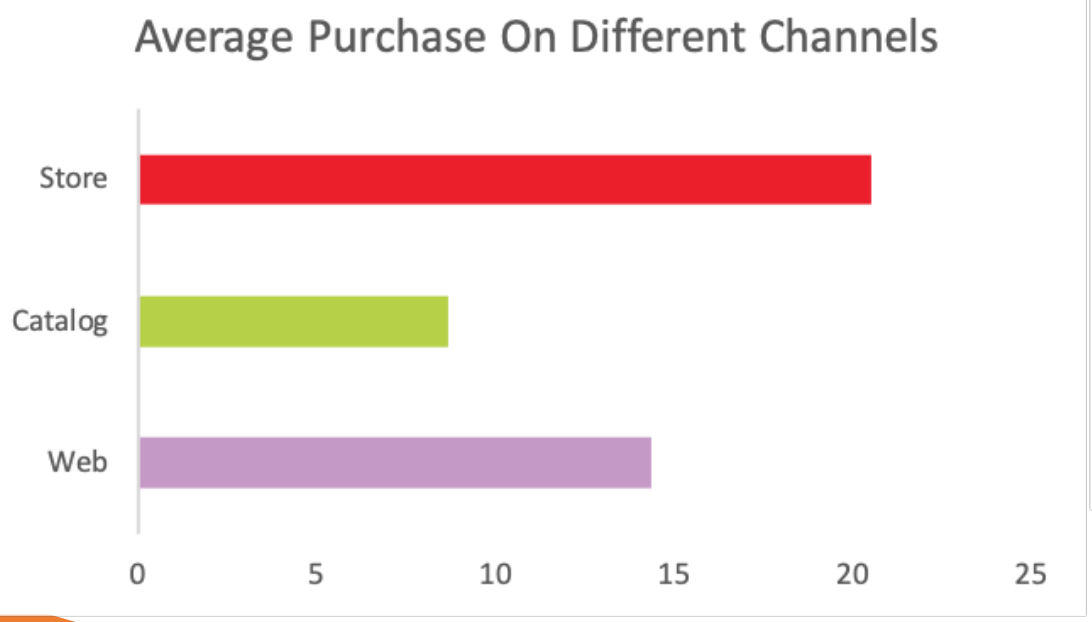
Master, Single PhD, Couple Without Kids/ Teens …

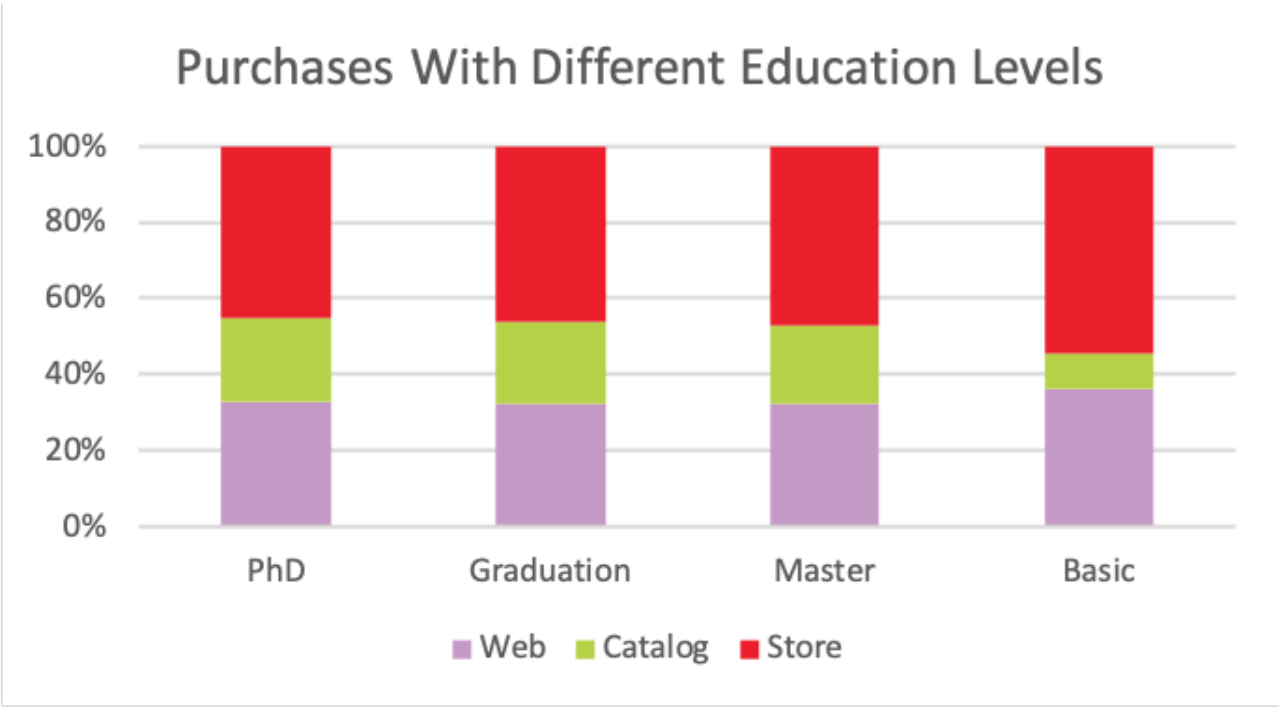
Assumption:
Customers with higher education levels and households without kids or teens are more willing to spend more on higher-priced groceries.

High Income, Consumer Hedonism…

How Education Level Can Influence Channel Choice?

Customers make more purchases directly in stores. There's no significant correlation between Education Level and Channel Choice.





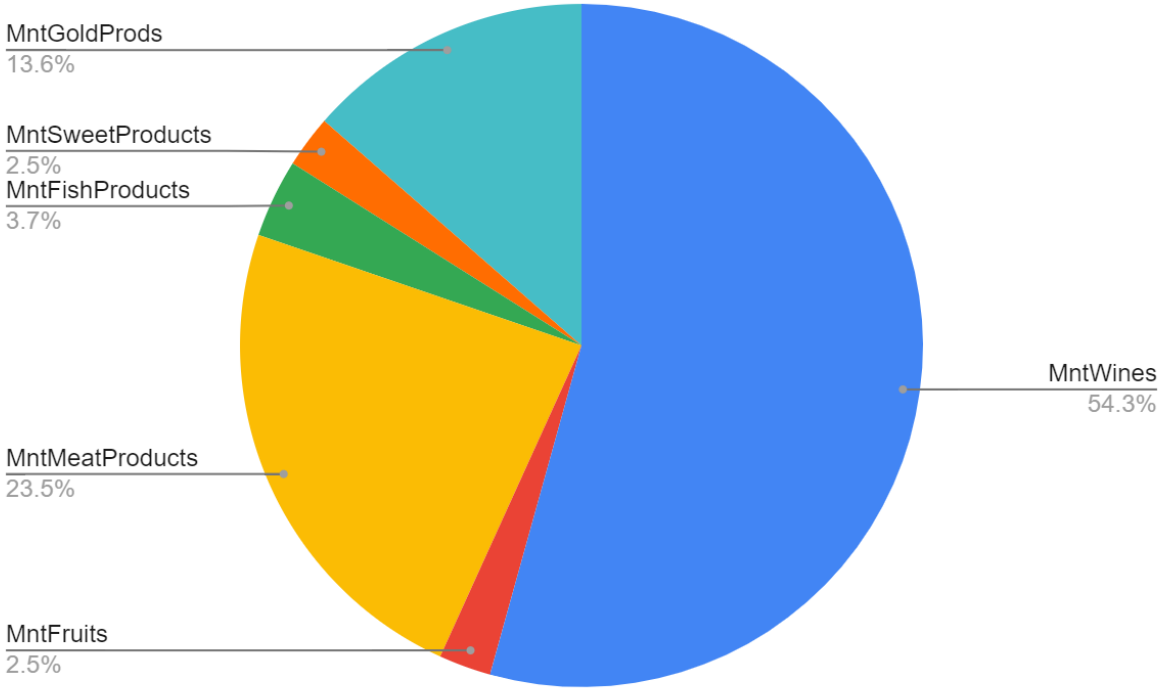
Assumption:
Households with kids or teens are more willing to make purchases online.

Does having a family affect buying decisions?

Amount Spent On Various Categories By Non-Parents vs Parents

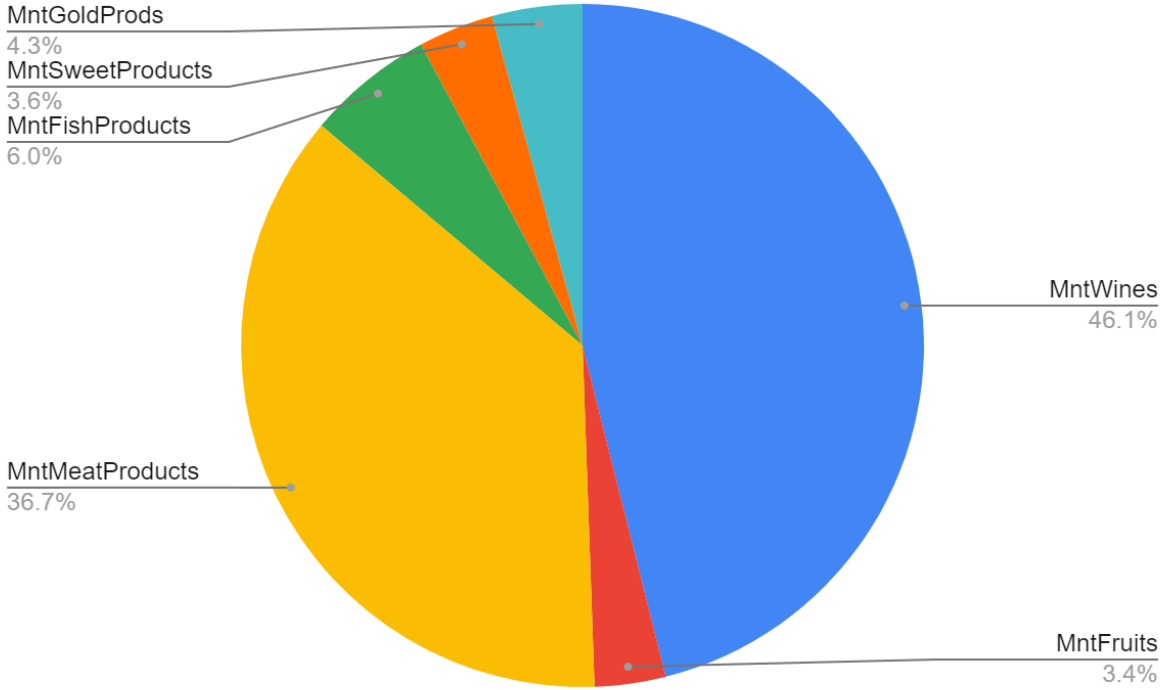
MntWines	MntFruits	MntMeat Products	MntFish Products	MntSweet Products	MntGold Prods
44	2	19	3	2	11

Parents
Count: 419



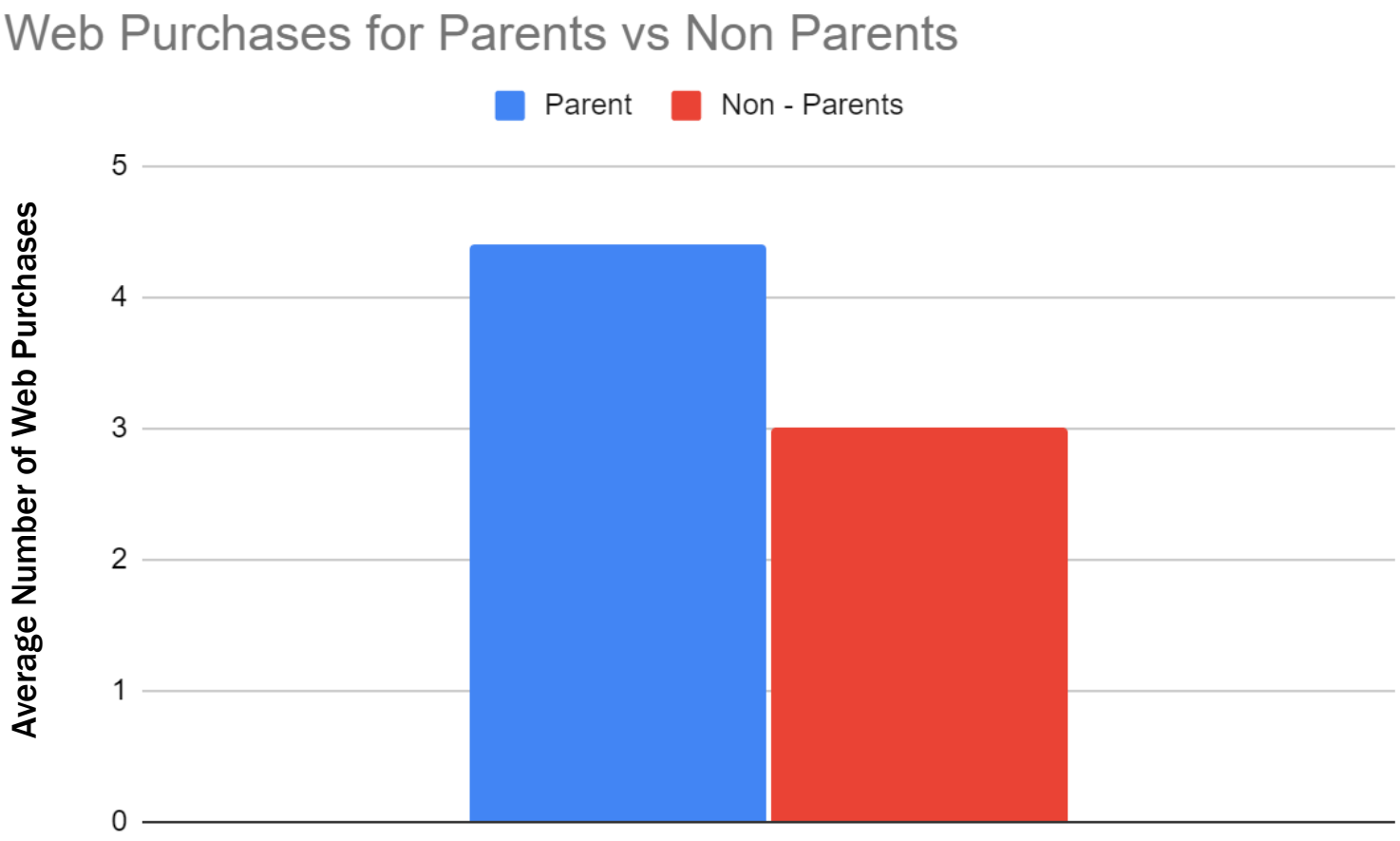
MntWines	MntFruits	MntMeatProducts	MntFish Products	MntSweet Products	MntGoldProds
452	33	360	59	35	42

Non-Parents
Count: 633

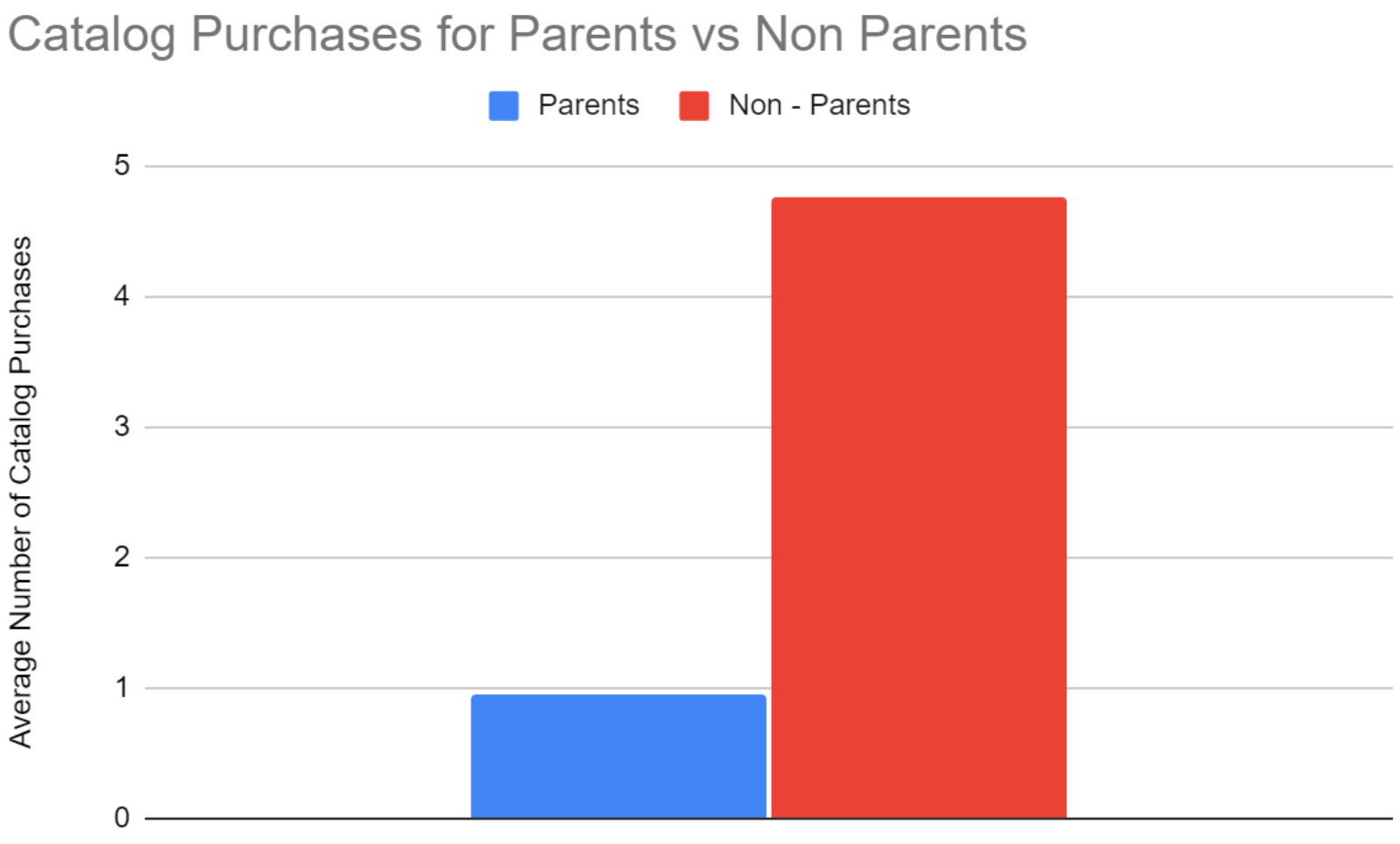


But Where Do They Shop From?

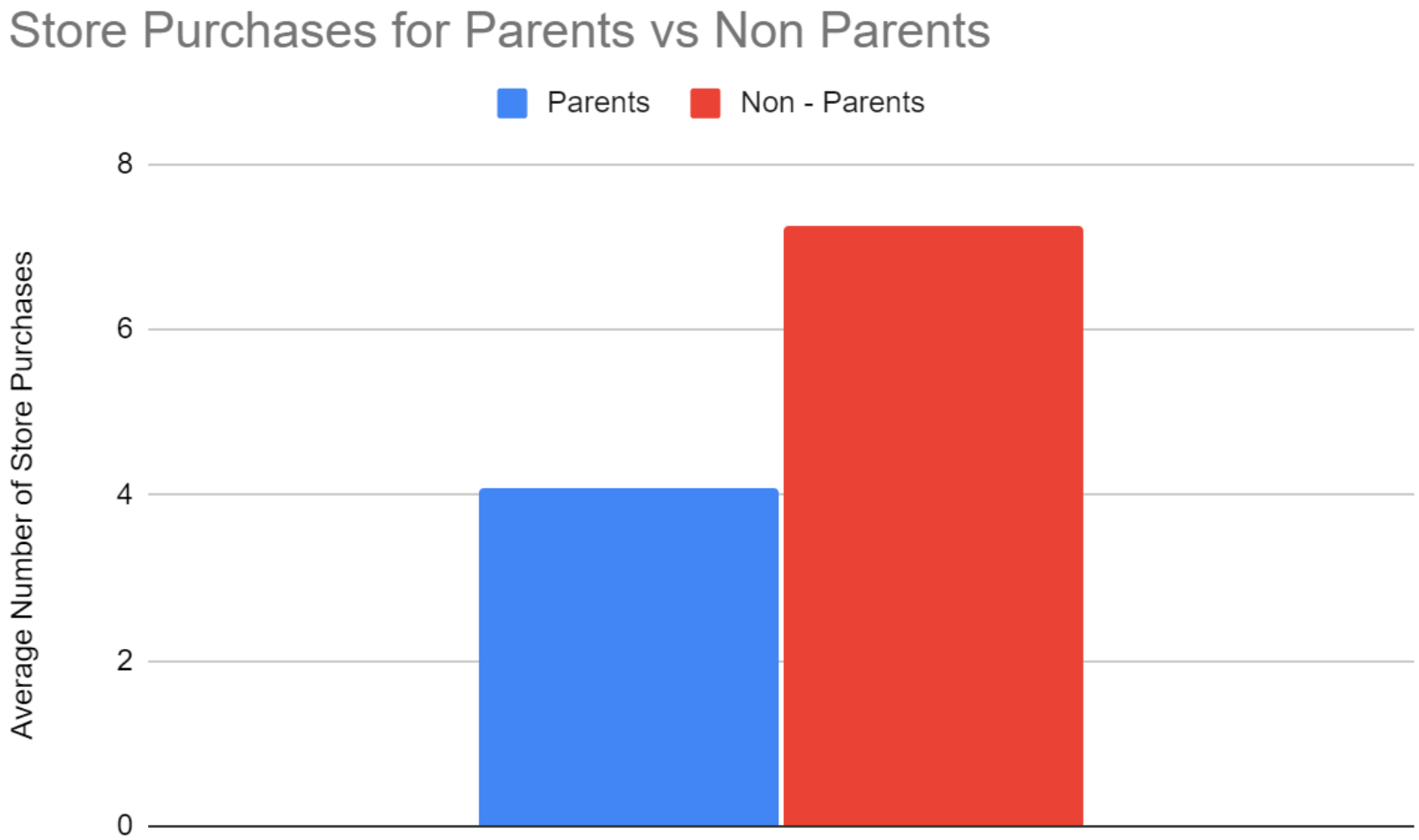
We tried to analyze if having a family also affects where one buys their products from.



But Where Do They Shop From?



But Where Do They Shop From?

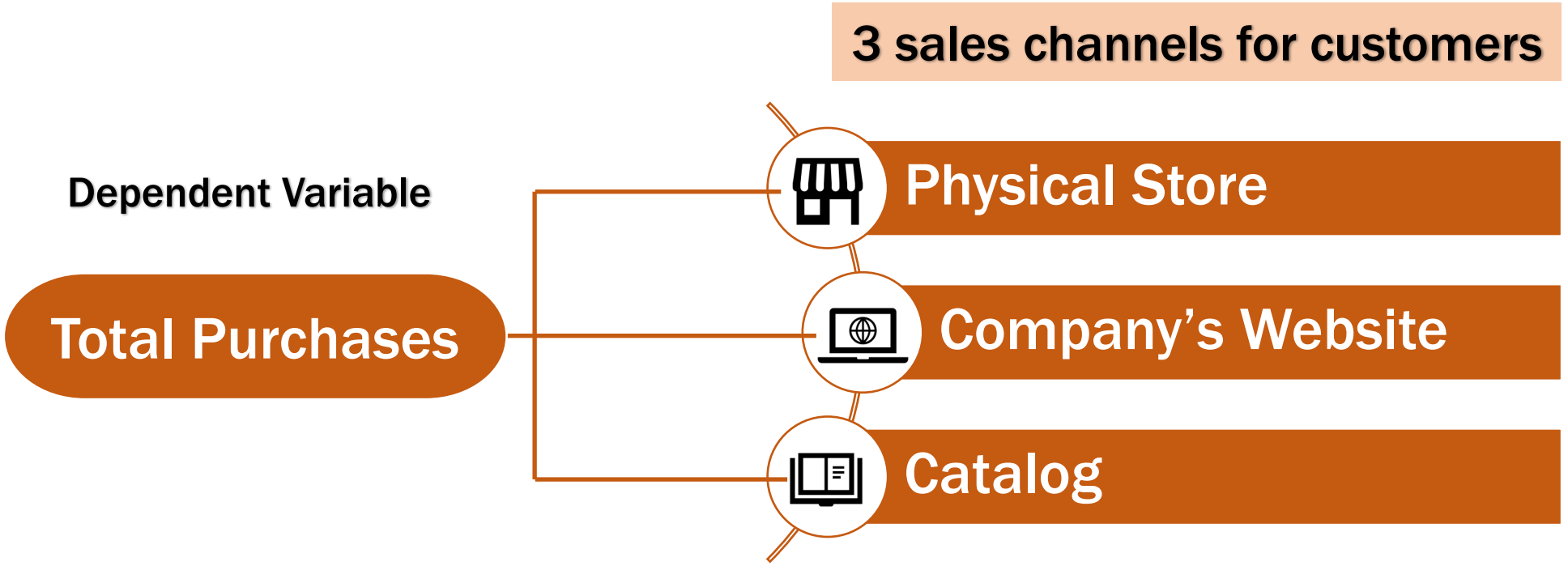


Final Thought

Having a family does not necessarily change what the client purchases, but it does affect which channel they choose to purchase from.

Predictive Models

Predictive Model A: No. of Purchases vs Sales Channel



Predictive Model A: No. of Purchases vs Sales Channel

Independent Variables:
30 variables

Stepwise Regression:

Stepwise regression performs selection based on statistical criteria to add or remove variables from the model until the best-fitting model is achieved.

Income	Response
Kidhome	Age
Teenhome	Customer_Days
Recency	marital_Divorced
MntWines	marital_Married
MntFruits	marital_Single
MntMeatProducts	marital_Together
MntFishProducts	marital_Widow
MntSweetProducts	education_2n Cycle
MntGoldProds	education_Basic
NumDealsPurchases	education_Graduation
NumWebVisitsMonth	education_Master
Complain	education_PhD
Z_CostContact	MntTotal
Z_Revenue	MntRegularProds

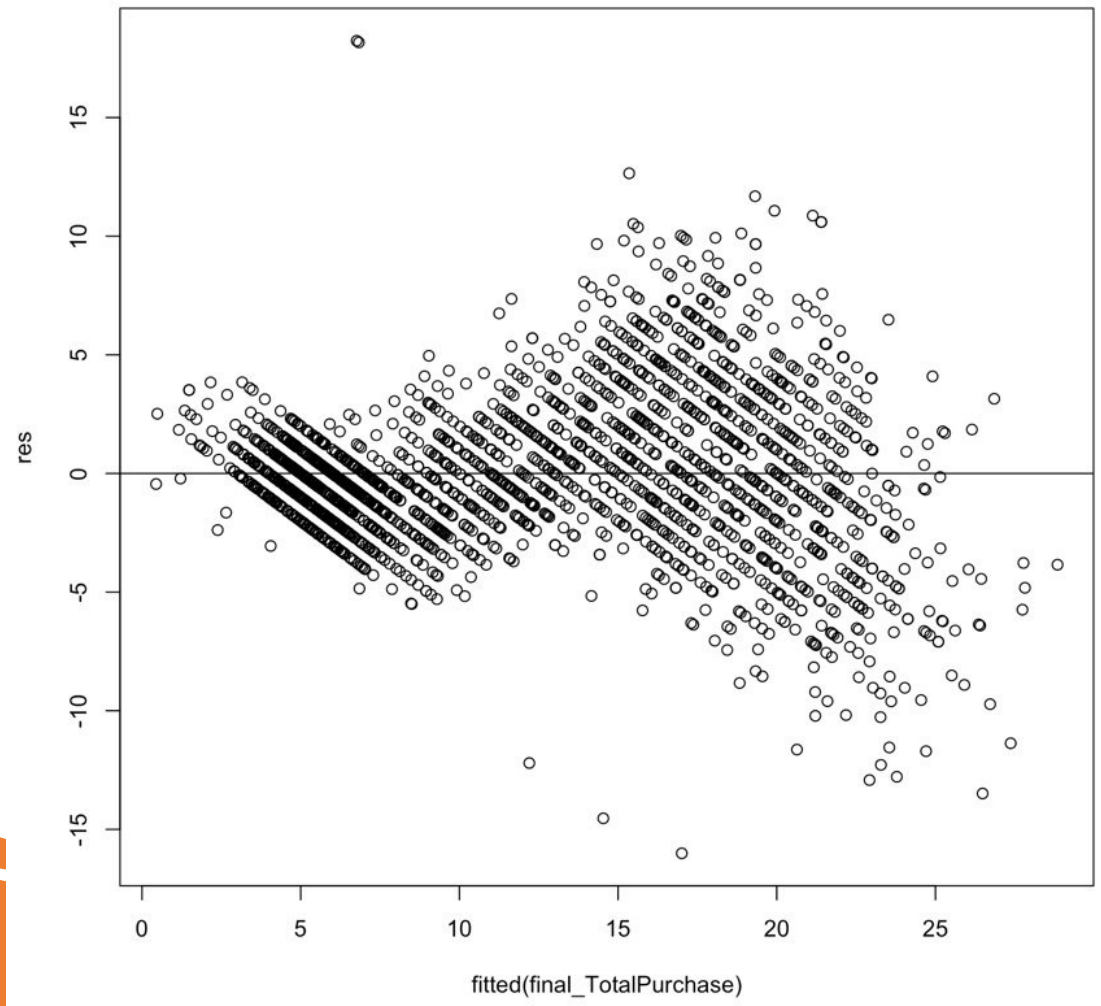
Predictive Model A: Stepwise Regression: 30 > 14 Variables

Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.0070	1.0450	-0.9640	0.3352	
Income	0.0001	0.0000	17.3720	< 2e-16	***
Kidhome	-2.3730	0.1721	-13.7870	< 2e-16	***
Teenhome	-0.4282	0.1654	-2.5890	0.0097	**
MntWines	0.0062	0.0003	17.9290	< 2e-16	***
MntFruits	0.0094	0.0025	3.7700	0.0002	***
MntMeatProducts	0.0014	0.0006	2.5380	0.0112	*
MntFishProducts	0.0072	0.0019	3.7950	0.0002	***
MntSweetProducts	0.0118	0.0024	4.9060	0.0000	***
MntGoldProds	0.0135	0.0017	8.1340	0.0000	***
NumDealsPurchases	0.8311	0.0457	18.2020	< 2e-16	***
Response	-0.8131	0.2138	-3.8030	0.0001	***
Customer_Days	0.0014	0.0004	3.5760	0.0004	***
education_Graduation	-0.2332	0.1637	-1.4250	0.1544	
education_Master	-0.3446	0.2162	-1.5940	0.1110	

- **Income, Kidhome, MntWines, NumDealsPurchases** are strongly significant
- **Education** is not statistically significant to the model

Legend on the significance level:
0 :***
0.001: **
0.01: *
0.05
0.1

Predictive Model A: Residual Plot & Result

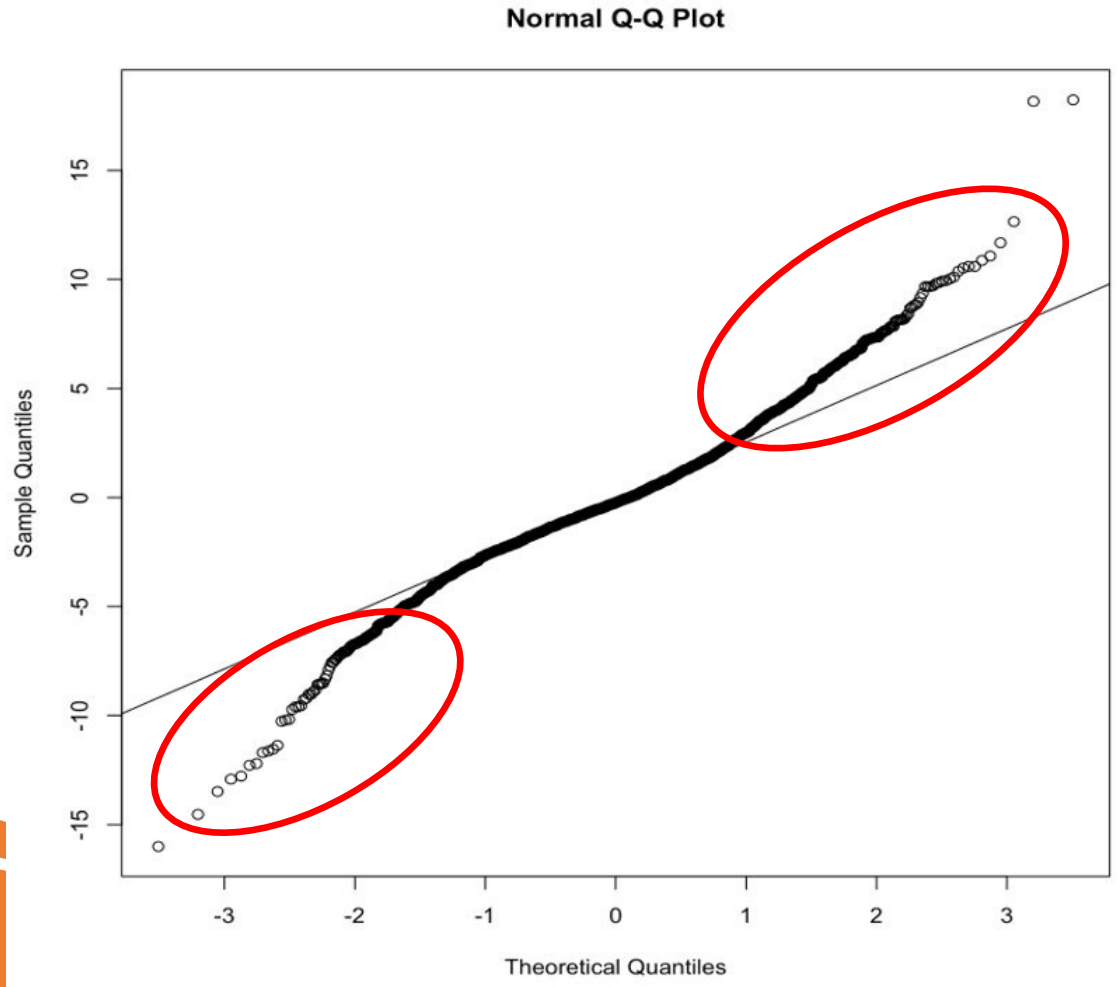


Residual Standard Error: 3.353
R-squared: 0.7833
Adjusted R-squared: 0.782

F-statistics: 565.6
P-value: $< 2 \times 10^{-16}$

- ✓ **78.33%** of the variability in the dependent variable is accounted for by the independent variables.
- ✓ The p-value is less than 0.0000 indicating that the model is **statistically significant**.

Predictive Model A: Normal QQ Plot



A normal Q-Q plot is a graphical tool used to assess whether a dataset follows a normal distribution.

- ✓ The largest values are larger than would be expected
- ✓ The smallest values are smaller than would be expected

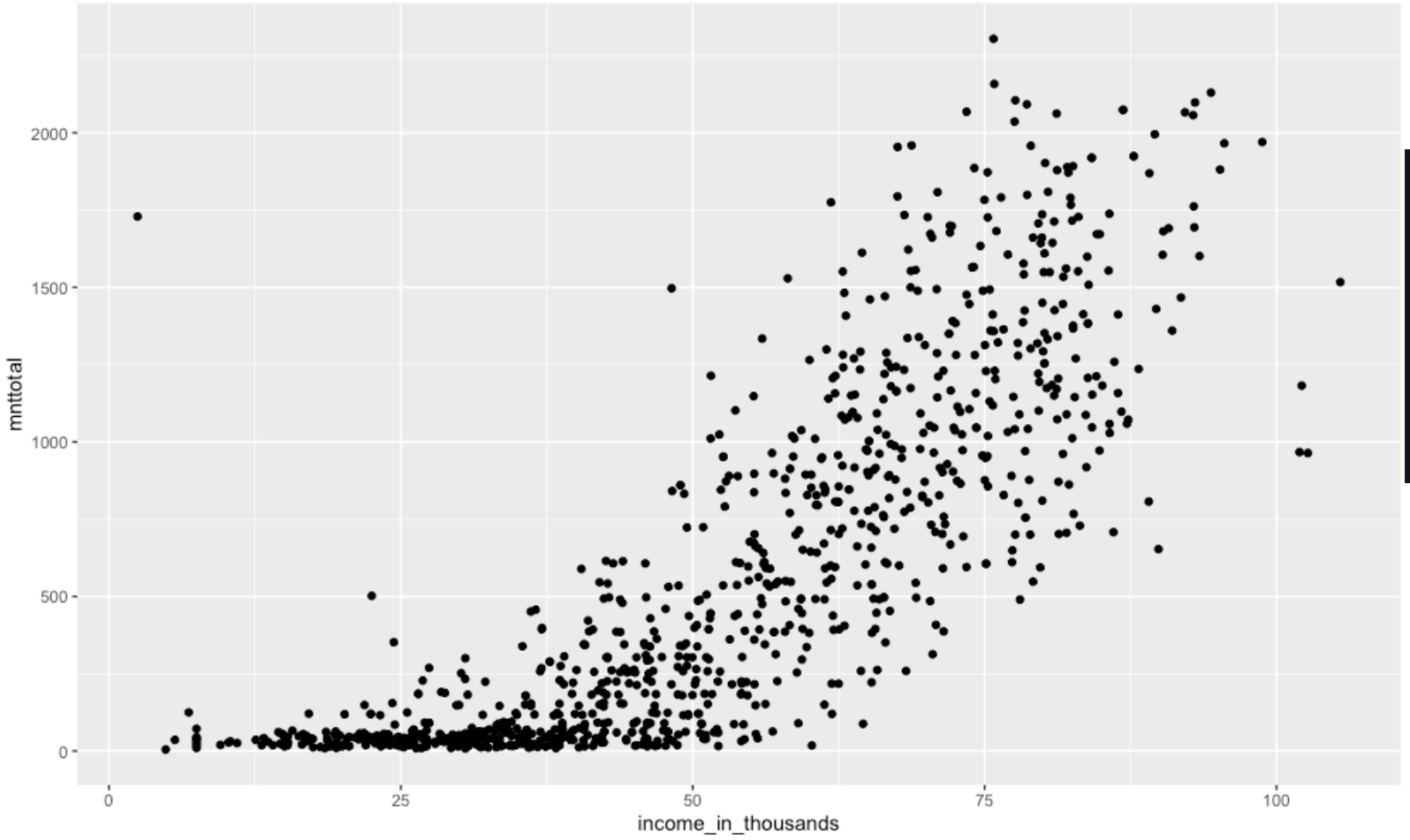
Predictive Model A: Amount Spent on Wines vs No. of Purchases

Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.0070	1.0450	-0.9640	0.3352	
MntWines	0.0062	0.0003	17.9290	< 2e-16	***

	# of Purchases below median	# of Purchases above median
Average spent on Wines	$63.5415	$ 553.0055

Customers spend more on wines, the more purchases they make!!

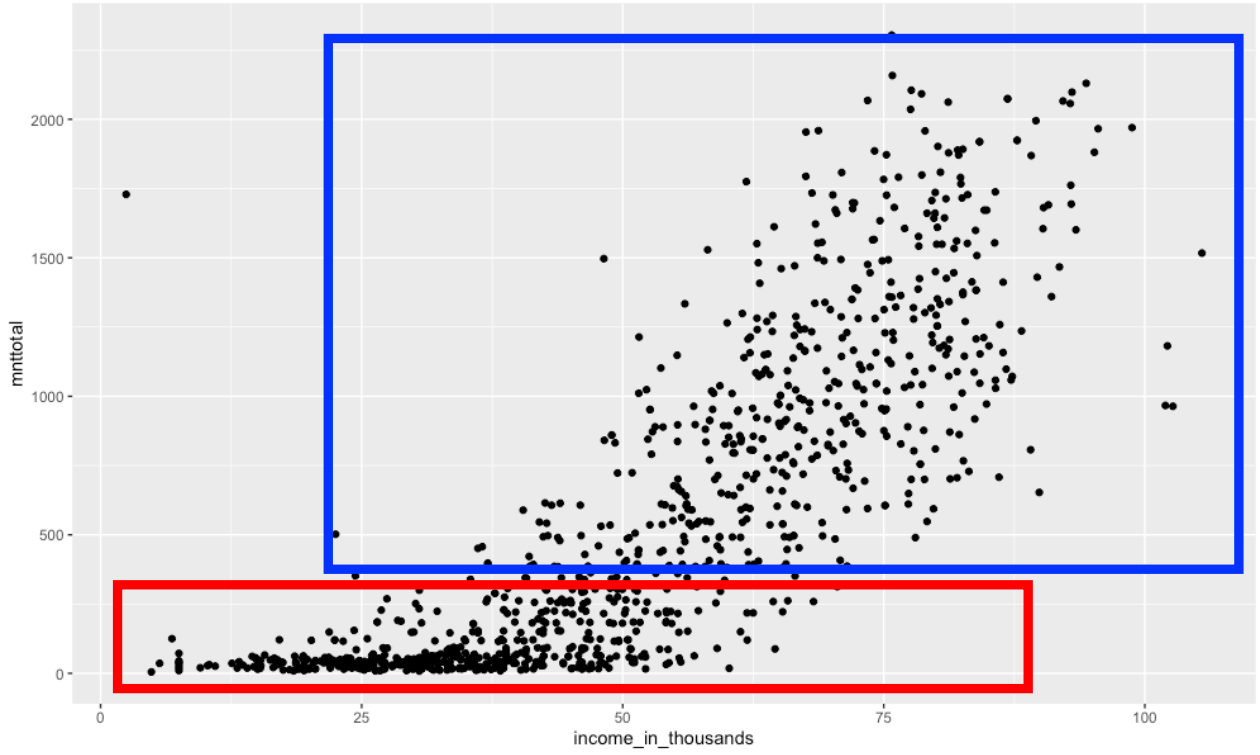
Predictive Model B: Income vs Total Spend (mnttotal)



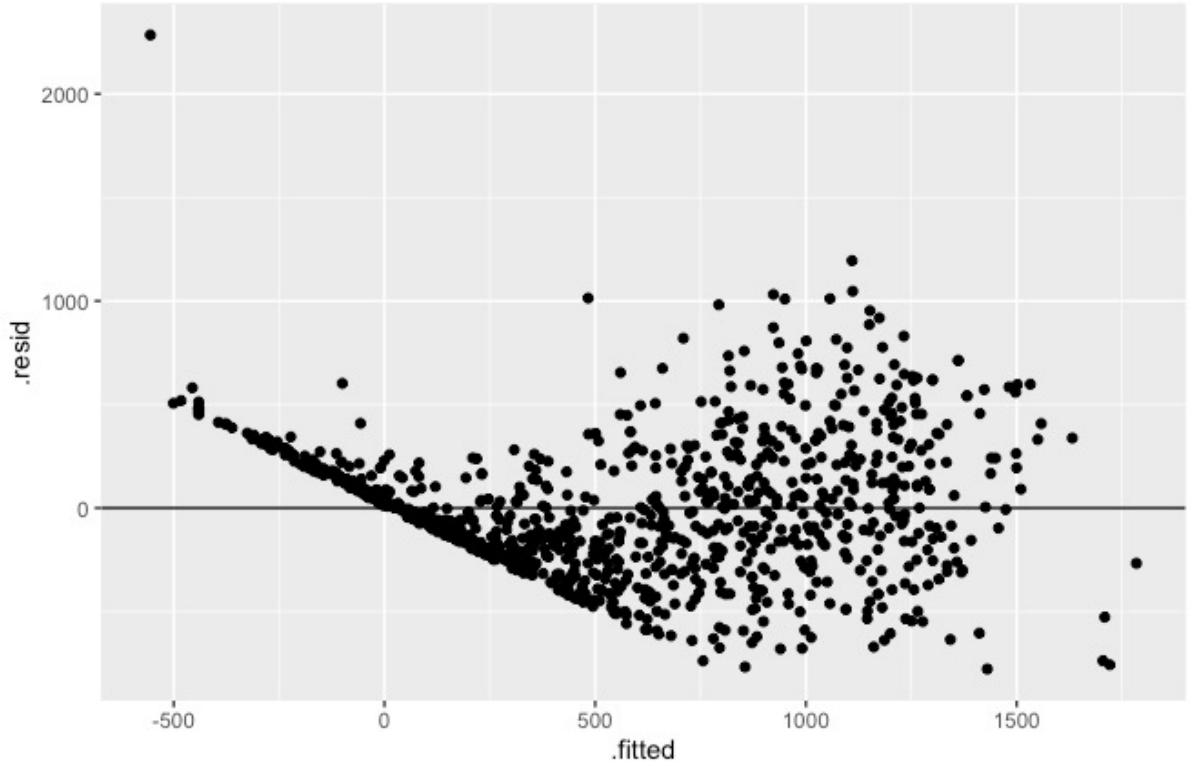

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -611.057     27.731  -22.04   <2e-16 ***
income_in_thousands   22.709      0.495   45.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 326.6 on 998 degrees of freedom
Multiple R-squared:  0.6783,    Adjusted R-squared:  0.678
F-statistic:  2104 on 1 and 998 DF,  p-value: < 2.2e-16
```

Predictive Model B: Two Issues

Two groups that behaved differently:



Heteroskedasticity (residuals are not normally distributed):

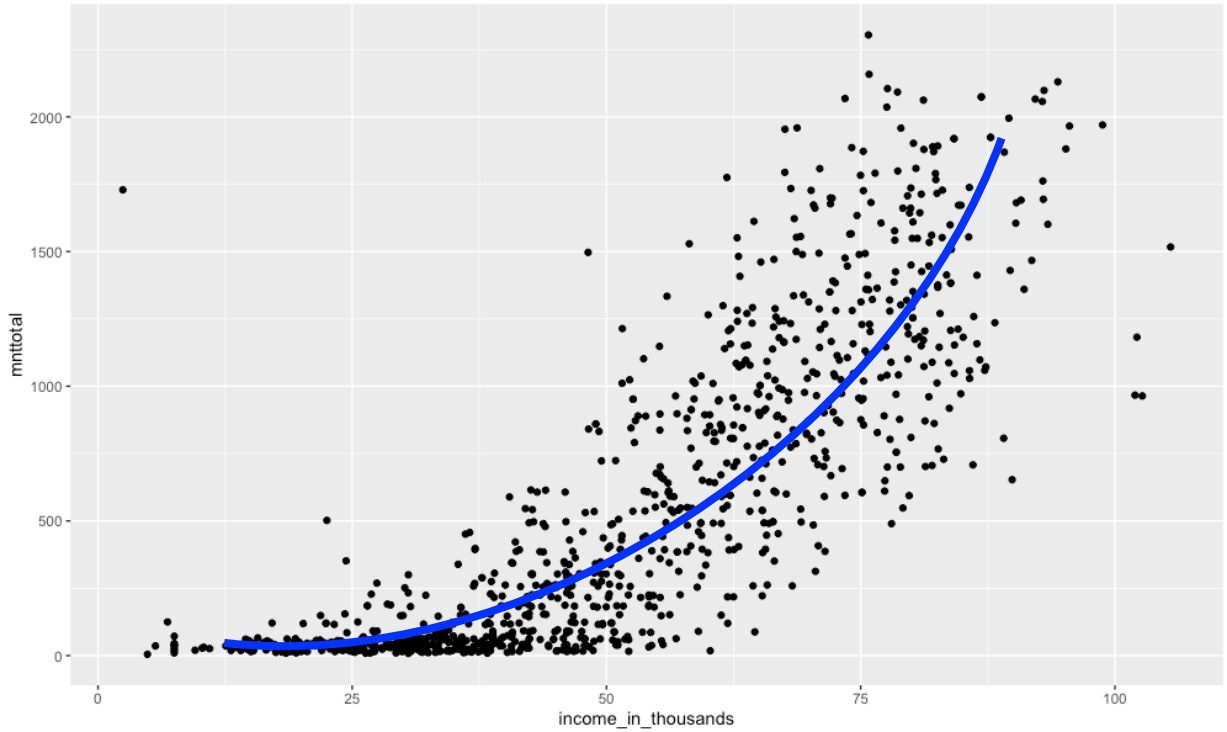


Make it rain

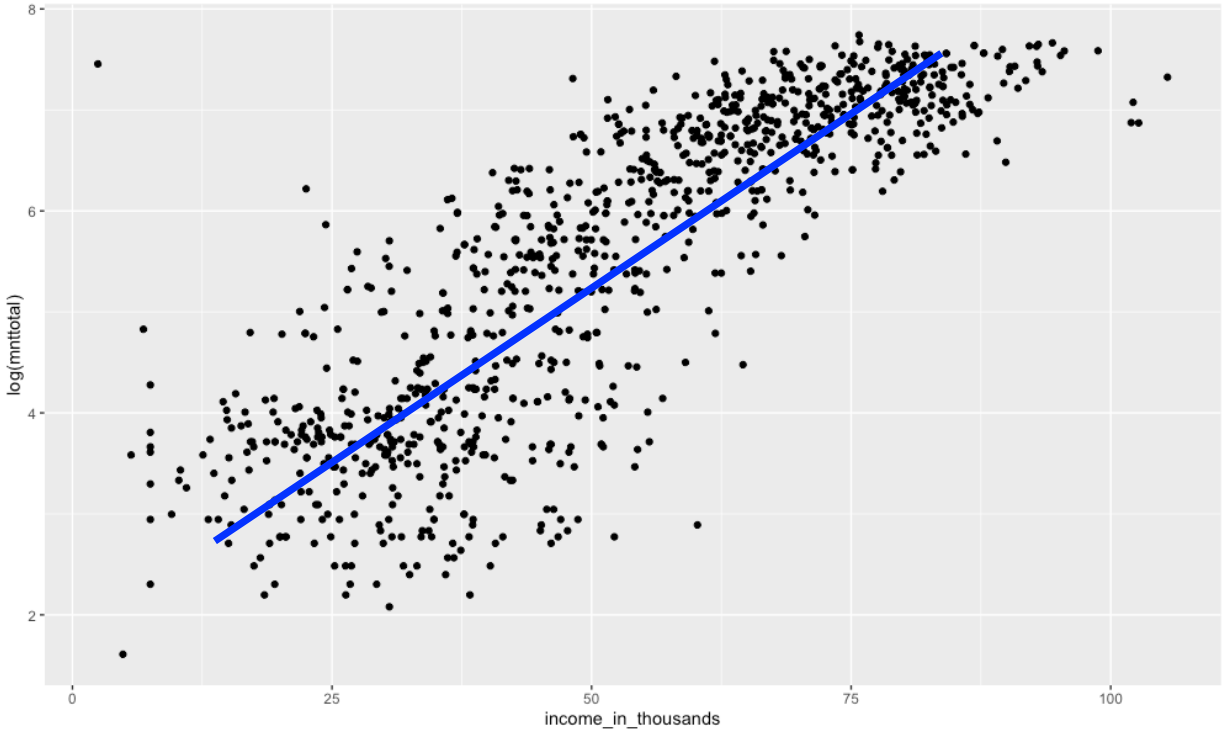
No money, no groceries

Predictive Model B: Solve for Heteroskedasticity

Raw (mnttotal):



Log (mnttotal):

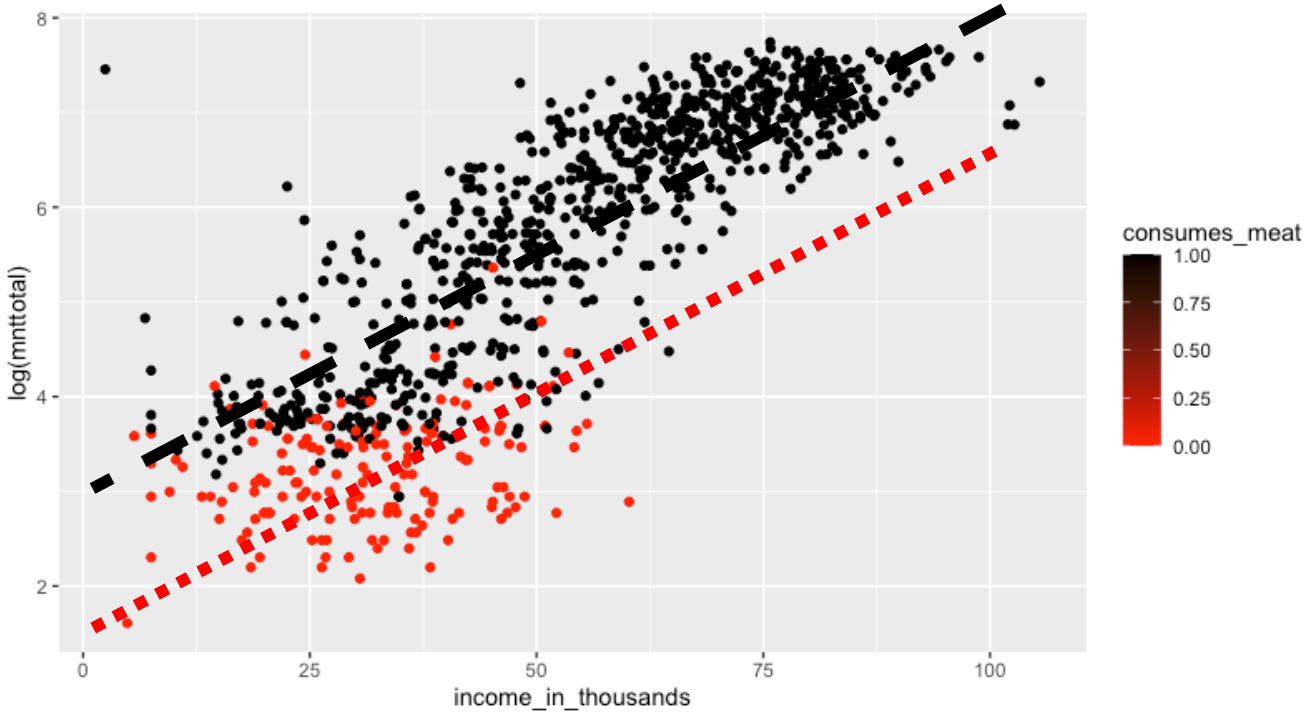


Predictive Model B: Solve for Differing Behavior

Guess this variable:

Predictive Model B: Final Model

Log(mnttotal) vs. Income and Consumes Meat

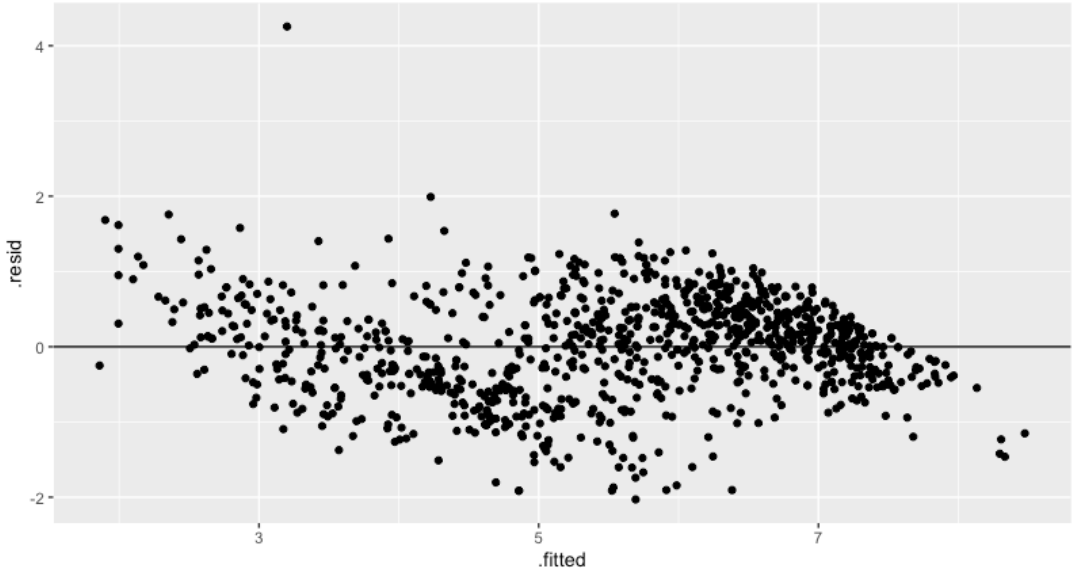


Slope Lines are approximated

Regression Output

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.611101   0.063181   25.50   <2e-16 ***
income_in_thousands 0.051203   0.001127   45.43   <2e-16 ***
consumes_meat      1.464287   0.063341   23.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.67 on 997 degrees of freedom
Multiple R-squared:  0.8126,    Adjusted R-squared:  0.8122
F-statistic:  2161 on 2 and 997 DF,  p-value: < 2.2e-16
```

Fitted vs. Residuals



Conclusion: Business Applications

- Acquire High-Income Customers >> More Purchases and LTV
- Meat and Wine consumption = Proxy for Income
  - Align campaigns to social and cultural norms that influence meat/wine consumption
- Family status and education level are not as predictive as spending power
- Stakeholders should target customers based on income
  - Rather than what they purchase or where they purchase

Questions?

Appendix A

```
-- Clean Data
#Remove the row with Income blank
#Remove the row with Marital Status "Absurd" and "Yolo"
#Change the marital status of the customer with "Alone" to "Single"
#Add a clumn of "Have_Kid_Or_Not", described as whether the customer has kids or teens at home
#Change the education level of the customer with "2n Cycle" to "Master"
create table ifood_data_cleaned_1
as
select *,
case
when Education = "2n Cycle" then "Master"
else Education
end as Education_clean
from
(select *,
case
when Kidhome = 0 and Teenhome = 0 then "No"
else "Yes"
end as Have_Kid_Or_Not
from
(select *,
case
when Marital_Status = "Alone" then "Single"
else Marital_Status
end as Marital_Status_clean
from
(select *,
case
when Income = '' then null
else Income
end as Income_clean
from ifood_data) as sub1) as sub2) as sub3
where Income_clean is not null and Marital_Status <> "YOLO" and Marital_Status <> "Absurd";
```

```
#Check if there're any duplicate IDs
select count(*) as total_count, count(distinct ID) as
count_remove_duplicate
from ifood_data_cleaned_1;
```

Appendix A

```
-- Explore the influence of Education Levels on Grocery Spending
#Overall picture
select Education_clean, avg_income, wines_spending + fruits_spending + meat_spending +
fish_spending + sweet_product + gold_spending as avg_total,
    wines_spending, fruits_spending, meat_spending, fish_spending, sweet_product,
gold_spending
from
(select Education_clean, avg(Income) as avg_income,
    avg(MntWines) as wines_spending, avg(MntFruits) as fruits_spending,
    avg(MntMeatProducts) as meat_spending, avg(MntFishProducts) as fish_spending,
    avg(MntSweetProducts) as sweet_product, avg(MntGoldProds) as gold_spending
from ifood_data_cleaned_1
group by Education_clean
order by avg_income desc) as sub
order by avg_total desc;

#Customer Profile for different products
##Wine (Top1 product having the most spending amount)
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(MntWines) as wines_spending
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by wines_spending desc
limit 5;
##Meat (Top2 product having the most spending amount)
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(MntMeatProducts) as meat_spending
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by meat_spending desc
limit 5;
##Gold
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(MntGoldProds) as gold_spending
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by gold_spending desc
limit 5;
##Fish
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(MntFishProducts) as fish_spending
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by fish_spending desc
limit 5;
##Sweet
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(MntSweetProducts) as sweet_spending
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by sweet_spending desc
limit 5;
##Fruit
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(MntFruits) as fruit_spending
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by fruit_spending desc
limit 5;
```

Appendix A

```
-- Explore the influence of Education Levels on choosing Channel Purchase
#Overall picture
select Education_clean, web, catalog, store,
    web + catalog + store as total, deal
from
(select Education_clean, avg(NumWebPurchases) as web, avg(NumCatalogPurchases) as catalog,
avg(NumStorePurchases) as store,
    avg(NumDealsPurchases) as deal
from ifood_data_cleaned_1
group by Education_clean) as sub
group by Education_clean
order by total desc;

#Customer Profile for different channels
##Web
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(NumWebPurchases) as web
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by web desc
limit 5;
##Store
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(NumStorePurchases) as store
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by store desc
limit 5;

##Catalog
select Education_clean, Marital_Status_clean, Have_Kid_Or_Not,
    avg(NumCatalogPurchases) as catalog
from ifood_data_cleaned_1
group by Education_clean, Marital_Status_clean, Have_Kid_Or_Not
order by catalog desc
limit 5;
```

Appendix B

```
#count of parents

select count(ID) as Parents
from ml_project1_data
where kidhome > 0 and teenhome > 0;

#count of non-parents
select count(ID) as Not_Parents
from ml_project1_data
where kidhome = 0 and teenhome = 0;

#amnt spent for parents for each category

select ID as Parents_ID, MntWines, MntFruits,
MntMeatProducts, MntFishProducts, MntSweetProducts,
MntGoldProds from ml_project1_data

where ID in (select ID from ml_project1_data

where kidhome > 0 and teenhome > 0 )

order by ID;

#amnt spent for non-parents for each category

select ID as Non_Parents_ID, MntWines, MntFruits,
MntMeatProducts, MntFishProducts, MntSweetProducts,
MntGoldProds from ml_project1_data

where ID in (select ID from ml_project1_data

where kidhome = 0 and teenhome = 0 )

order by ID;
```

Appendix B

```
#WEB - Parents

select ID as Parents, NumWebPurchases from
ml_project1_data

where ID in (select ID from ml_project1_data

where kidhome > 0 and teenhome > 0 )

order by ID;

###WEB - Non - Parents

select ID as Non_Parents, NumWebPurchases from
ml_project1_data

where ID in (select ID from ml_project1_data

where kidhome = 0 and teenhome = 0 )

order by ID;

#Catalog - Parents

select ID as Parents, NumCatalogPurchases from
ml_project1_data

where ID in (select ID from ml_project1_data

where kidhome > 0 and teenhome > 0 )

order by ID;

#Catalog - Non -Parents

select ID as Non_Parents, NumCatalogPurchases from
ml_project1_data

where ID in (select ID from ml_project1_data

where kidhome = 0 and teenhome = 0 )

order by ID;

#Store Purchases - Parents

select ID as Parents, NumStorePurchases, Income from
ml_project1_data

where ID in (select ID from ml_project1_data

where kidhome > 0 and teenhome > 0 )

order by ID;

#Store Purchases - Non - Parents

select ID as Non_Parents, NumStorePurchases from
ml_project1_data

where ID in (select ID from ml_project1_data

where kidhome = 0 and teenhome = 0 )

order by ID;
```

Appendix C

```
##R code for performing stepwise regression##
ifood <- read.csv("/Users/chaomanling/Desktop/MSBA/Summer/Data Management and
SQL/ifood_df.csv", header=T)
head(ifood)

###Total number of purchase #R=0.7833
ifood$sum_purchase <-
 ifood$NumStorePurchases + ifood$NumWebPurchases + ifood$NumCatalogPurchases
x4 <- colnames(ifood[, -c(12:14, 16:20, 39, 40)])
formula_str4 <- paste("sum_purchase~", paste(x4, collapse = " + "))
TotalPurchase <- lm(formula_str4, data = ifood)
final_TotalPurchase <- step(TotalPurchase, direction = "both")
summary(final_TotalPurchase)

#residula plot
res <- resid(final_TotalPurchase)
plot(fitted(final_TotalPurchase), res)
abline(0,0)

#create Q-Q plot for residuals
qqnorm(res)
#add a straight diagonal line to the plot
qqline(res)
```

Appendix C

##My SQL Code##

```
ALTER TABLE ifood ADD sum_purchase INT;

UPDATE ifood
SET sum_purchase = NumWebPurchases + NumCatalogPurchases + NumStorePurchas
es;

#Compute the median of total purchase
SELECT AVG(sum_purchase) AS Median
FROM (
  SELECT sum_purchase, @rowindex:=@rowindex + 1 AS rowindex,
@total_rows AS total_rows
  FROM ifood, (SELECT @rowindex := 0, @total_rows := (SELECT COUNT(*)
FROM ifood)) r
  ORDER BY sum_purchase
) AS d
WHERE rowindex IN (FLOOR(total_rows / 2), CEIL(total_rows / 2));
```

Appendix C

```
#Compute the average amount spend on wines that purchases below the median
SELECT AVG(MntWines) AS avg_wines_lower
FROM ifood
WHERE sum_purchase <= (
   SELECT AVG(sum_purchase) AS Median
   FROM (
      SELECT sum_purchase, @rowindex:=@rowindex + 1 AS rowindex,
@total_rows AS total_rows
      FROM ifood, (SELECT @rowindex := 0, @total_rows := (SELECT COUNT(*)
FROM ifood)) r
      ORDER BY sum_purchase
   ) AS d
   WHERE rowindex IN (FLOOR(total_rows / 2), CEIL(total_rows / 2))
);
#Compute the average amount spend on wines that purchases above the median
SELECT AVG(MntWines) AS avg_wines_upper
FROM ifood
WHERE sum_purchase > (
   SELECT AVG(sum_purchase) AS Median
   FROM (
      SELECT sum_purchase, @rowindex:=@rowindex + 1 AS rowindex,
@total_rows AS total_rows
      FROM ifood, (SELECT @rowindex := 0, @total_rows := (SELECT COUNT(*)
FROM ifood)) r
      ORDER BY sum_purchase
   ) AS d
   WHERE rowindex IN (FLOOR(total_rows / 2), CEIL(total_rows / 2))
);
```

Appendix D

SQL script to clean data for regression analysis:

```
#DROP TABLE ifood_clean;

CREATE TABLE ifood_clean as (
SELECT
  income
, income / 1000 as income_in_thousands
, kidhome
, teenhome
, recency
, mntwines
, CASE
    WHEN mntwines > 10 THEN 1
    ELSE 0
  END as consumes_wine
, mntfruits
, CASE
    WHEN mntfruits > 10 THEN 1
    ELSE 0
  END as consumes_fruits
, mntmeatproducts
, CASE
    WHEN mntmeatproducts > 10 THEN 1
    ELSE 0
  END as consumes_meat
, mntfishproducts
, CASE
    WHEN mntfishproducts > 10 THEN 1
    ELSE 0
  END as consumes_fish
, mntsweetproducts
, CASE
    WHEN mntsweetproducts > 10 THEN 1
    ELSE 0
  END as consumes_sweets
, mntgoldprods
, CASE
    WHEN mntgoldprods > 10 THEN 1
    ELSE 0
  END as consumes_gold
, NumDealsPurchases as num_deals_purchases
, NumWebPurchases as num_web_purchases
, NumCatalogPurchases as num_catalog_purchases
, NumStorePurchases as num_store_purchases
, NumWebVisitsMonth as num_web_visits_month
, AcceptedCmp1 as accepted_cmp1
, AcceptedCmp1 as accepted_cmp2
, AcceptedCmp1 as accepted_cmp3
, AcceptedCmp1 as accepted_cmp4
, AcceptedCmp1 as accepted_cmp5
, complain , Z_CostContact as z_cost_contact
, z_revenue
, response
, age
, customer_days
, marital_divorced
, marital_single
, marital_together
, marital_widow
, 'education_2n Cycle' as education_2n_cycle
, education_basic
, education_graduation
, education_master
, education_phd
, mnttotal
, mntregularprods
, AcceptedCmpOverall as accepted_cmp_overall
FROM ifood_df_full
)
```

Appendix D

R code to create linear models and plots:

```
#install.packages("tidyverse")
#install.packages("SciViews")
library(ggplot2)
library("SciViews")
```

```
#regression, no log no meat
lm0 = lm(mnttotal~income_in_thousands,data=ifood_clean)
summary(lm0)
res = resid(lm0)
plot(fitted(lm0), res)
ggplot(data=lm0,aes(x=.fitted,y=.resid)) + geom_point() +
geom_hline(yintercept = 0)
```

```
#regression, no log on mnttotal
lm1 =
lm(mnttotal~income_in_thousands+consumes_meat,data=ifoo
d_clean)
summary(lm1)
res = resid(lm1)
plot(fitted(lm1), res)
```

```
#regression, log on mnttotal
lmlog = lm(log(mnttotal)~income_in_thousands+consumes_meat,data=ifood_clean)
summary(lmlog)
res = resid(lmlog)
plot(fitted(lmlog), res)
ggplot(data=lmlog,aes(x=.fitted,y=.resid)) + geom_point() + geom_hline(yintercept = 0)
```

```
#no log on income
plot_no_color = ggplot(data=ifood_clean,
aes(x=income_in_thousands,y=mnttotal))+geom_point()
plot_no_color
plot = ggplot(data=ifood_clean,
aes(x=income_in_thousands,y=mnttotal,color=consumes_meat))+geom_point()
plot +geom_point(size=0.1)+scale_color_gradient(low = "red", high = "black")
```

```
#log income
plot_no_color2 = ggplot(data=ifood_clean,
aes(x=income_in_thousands,y=log(mnttotal)))+geom_point()
plot_no_color2
plot2 = ggplot(data=ifood_clean,
aes(x=income_in_thousands,y=log(mnttotal),color=consumes_meat))
plot2 +geom_point()+scale_color_gradient(low = "red", high = "black")
```

```
#hist of meat consumption
ggplot(data=ifood_clean, aes(x=mntmeatproducts,color=income))+geom_histogram()
```