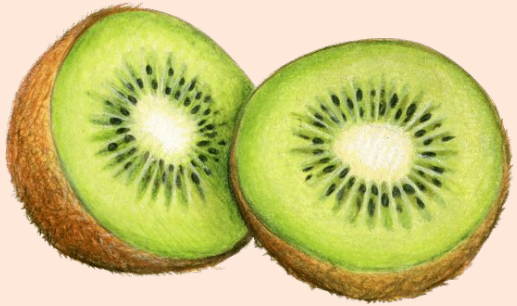
06/02/2022

Customer Personality Traits in Purchasing Fruits

Data Miners: Felix Gao, Leslie Vazquez Moreno, Sean Nguyen, Stella Aurelia, and Victoria Nguyen



Team Members & Roles

Felix Gao:
Exploring levels of income, Creating logistic regression model

Sean Nguyen:
Exploring marital status

Stella Aurelia:
Exploring levels of education

Leslie Vazquez Moreno:
Exploring age

Victoria Nguyen:
Exploring number of kids at customer's household

Agenda

- Dataset
- Project Objectives
- Recap + Subquestions
- Main Research Question
- Logistic Regression Model

The Dataset

This dataset is an analysis of a company's ideal customers designed to help businesses better understand its customers and cater their products to the customer's needs, behaviors and concerns.

This can be found at **kaggle.com**.

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

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits
5524	1957	Graduation	Single	58138	0	0	04-09-2012	58	635	88
2174	1954	Graduation	Single	46344	1	1	08-03-2014	38	11	1
4141	1965	Graduation	Together	71613	0	0	21-08-2013	26	426	49
6182	1984	Graduation	Together	26646	1	0	10-02-2014	26	11	4
5324	1981	PhD	Married	58293	1	0	19-01-2014	94	173	43
7446	1967	Master	Together	62513	0	1	09-09-2013	16	520	42
965	1971	Graduation	Divorced	55635	0	1	13-11-2012	34	235	65
6177	1985	PhD	Married	33454	1	0	08-05-2013	32	76	10
4855	1974	PhD	Together	30351	1	0	06-06-2013	19	14	0
5899	1950	PhD	Together	5648	1	1	13-03-2014	68	28	0
1994	1983	Graduation	Married		1	0	15-11-2013	11	5	5
387	1976	Basic	Married	7500	0	0	13-11-2012	59	6	16
2125	1959	Graduation	Divorced	63033	0	0	15-11-2013	82	194	61
8180	1952	Master	Divorced	59354	1	1	15-11-2013	53	233	2
2569	1987	Graduation	Married	17323	0	0	10-10-2012	38	3	14
2114	1946	PhD	Single	82800	0	0	24-11-2012	23	1006	22
9736	1980	Graduation	Married	41850	1	1	24-12-2012	51	53	5
4939	1946	Graduation	Together	37760	0	0	31-08-2012	20	84	5
6565	1949	Master	Married	76995	0	1	28-03-2013	91	1012	80
2278	1985	2n Cycle	Single	33812	1	0	03-11-2012	86	4	17
9360	1982	Graduation	Married	37040	0	0	08-08-2012	41	86	2
5376	1979	Graduation	Married	2447	1	0	06-01-2013	42	1	1
1993	1949	PhD	Married	58607	0	1	23-12-2012	63	867	
4047	1954	PhD	Married	65324	0	1	11-01-2014	0		
1409	1951	Graduation	Together	40689	0	1	18-03-2013	69		3
7892	1969	Graduation	Single	18589	0	0	02-01-2013	89		
2404	1976	Graduation	Married	53359	1	1	27-05-2013	4		

The Dataset

We decided to look specifically at the characteristics of customers who spend on fruits.

- Independent Variables:

  Year_Birth, Education, Marital_Status, Income, KidHome

- Dependent Variable:

  MntFruits

Variable_Name	Description
Year_birth	Customer's Birth Year
Education	Customer's Education Level
Marital_status	Customer's Marital Status
Income	Customer's Yearly Household Income
Kidhome	Number Of Children In Customer's Household
Mntfruits	Amount Spent On Fruits In Last 2 Years

Project Objectives

What characteristics makes someone more likely to purchase fruits?

- Does the level of education have an effect in purchasing more fruits?
- What age group purchased the most fruit on average?
- Are married people more likely to spend more on fruits?
- Is there a relationship between income and the amount spent on fruits?
- Does the number of kids one has at home affect how much they spend on fruit?

Subquestion #1:

Does the level of education have an effect in purchasing more fruits?

Does the level of education have an effect in purchasing more fruits?

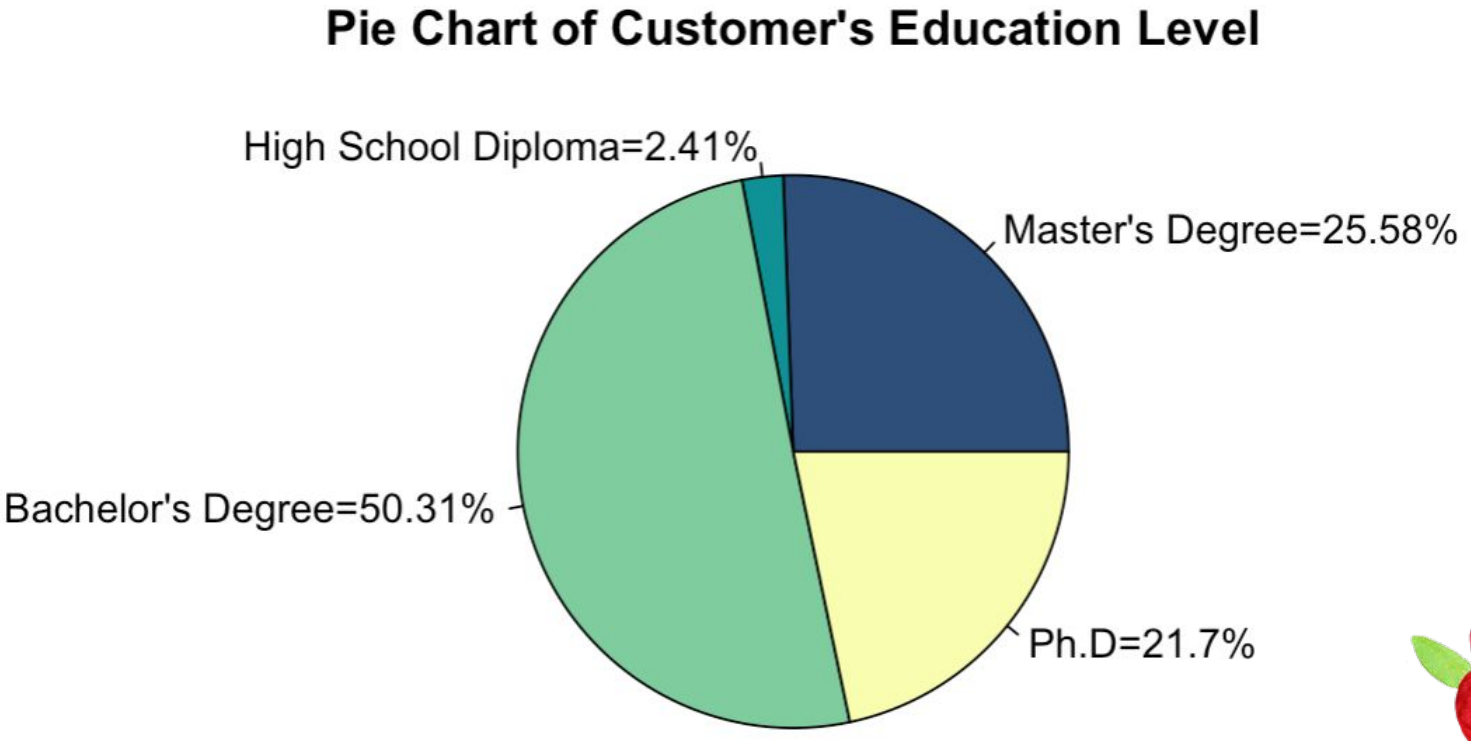
“Education” originally has 5 categories: 2n Cycle, Basic, Graduation, Master, PhD

```
marketing_campaign <-
marketing_campaign %>%
  mutate(Education = recode(Education,
  "2n Cycle" = "Master's Degree",
  "Basic" = "High School Diploma",
  "Master" = "Master's Degree",
  "Graduation" = "Bachelor's Degree",
  "PhD" = "Ph.D"))
```

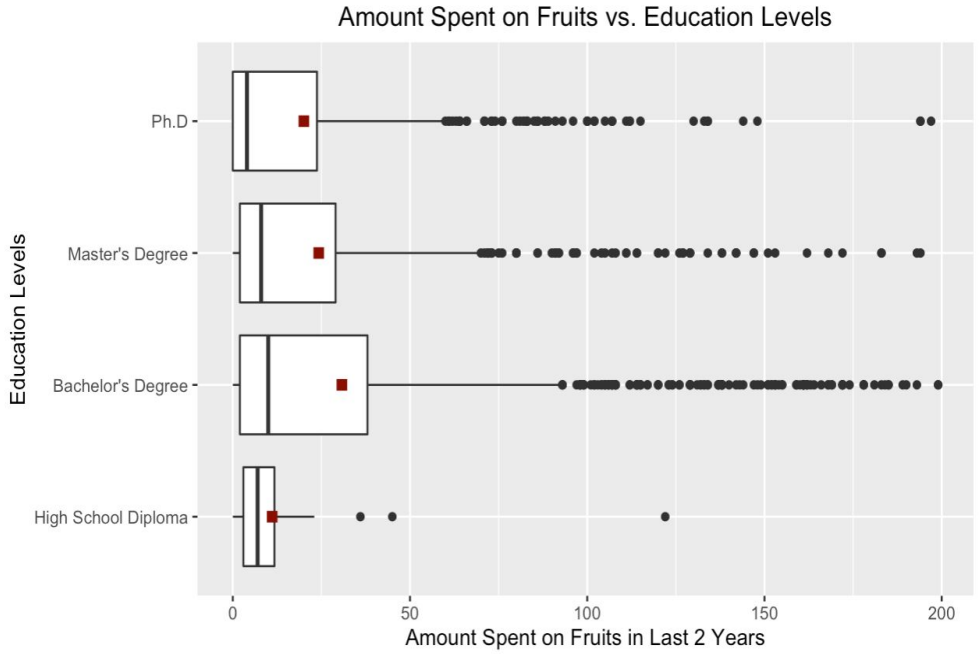
Recoded variables to 4 distinctive categories:

- High School Diploma
- Bachelor’s Degree
- Master’s Degree
- Ph.D

Does the level of education have an effect in purchasing more fruits?



Does the level of education have an effect in purchasing more fruits?



On average, those with a:

- High School Diploma spend **$11.11**
- Bachelor's Degree spend **$30.77**
- Master's Degree spend **$24.24**
- Ph.D spend **$20.04**

on fruits in the last two years.

■ Represents the mean of amount spent

Does the level of education have an effect in purchasing more fruits?

Chi-Squared Test

H_0: The two variables are independent.
H_a: The two variables relate to each other.

```
        Pearson's Chi-squared test

data:  MntFruits and Education
X-squared = 537.04, df = 471, p-value = 0.01876
```

Kruskal Wallis Test

H_0: There is no difference in the distribution of the populations
H_a: There is a difference in the distribution of the populations

```
        Kruskal-Wallis rank sum test

data:  MntFruits by Education
Kruskal-Wallis chi-squared = 47.578, df = 3, p-value = 2.619e-10
```

It appears that there is a relationship between **Education** and **MntFruits**. There is also a significant difference between education levels. Yes, the level of education has an effect in purchasing more fruits.



Subquestion #2:

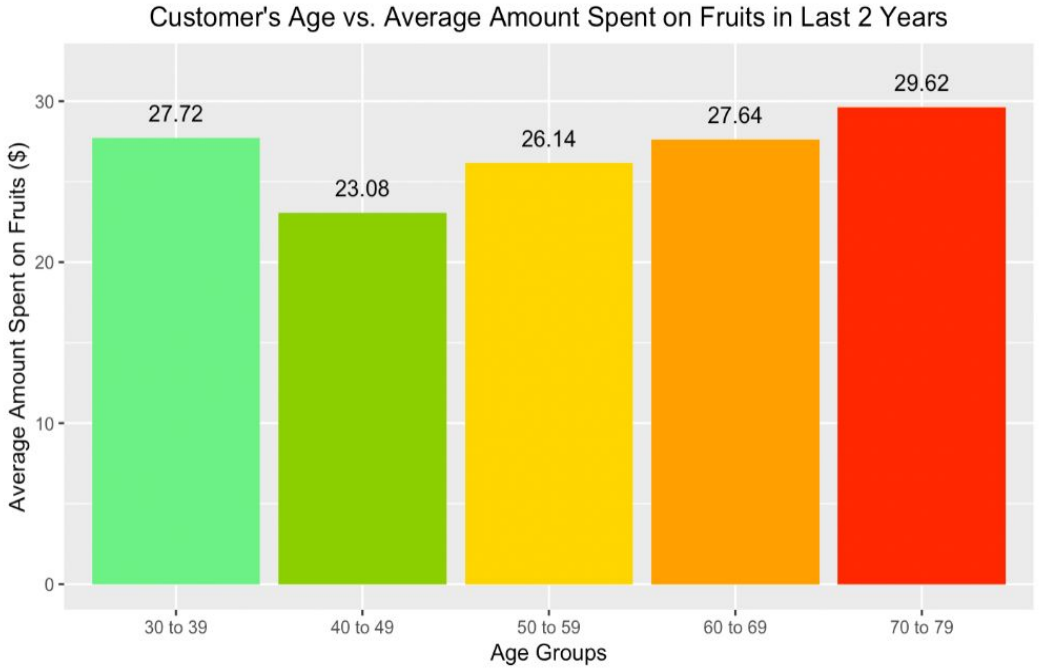
What age group purchased the most fruit on average?



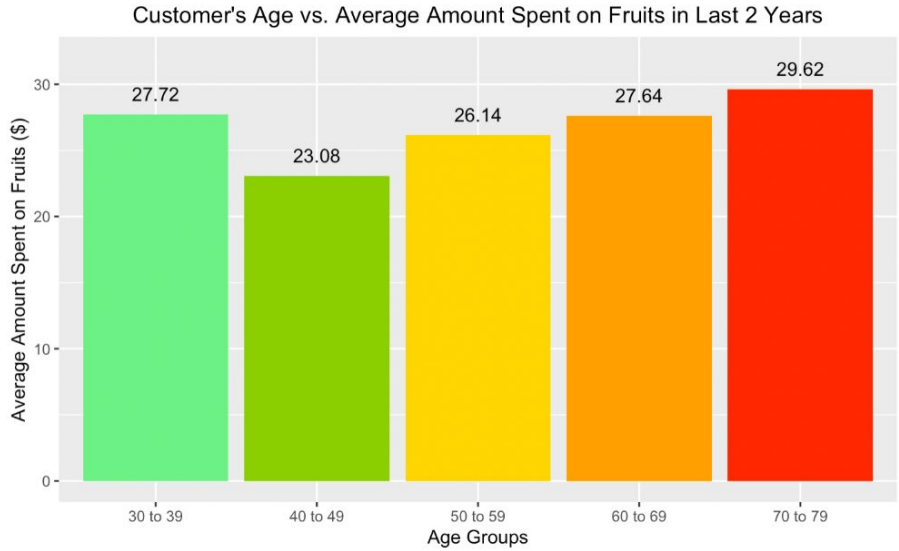
What age group purchased the most fruit on average?

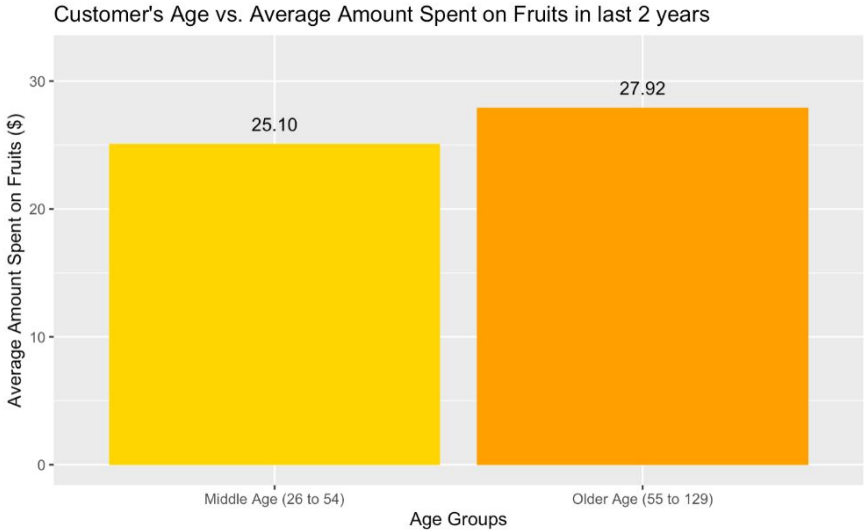
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
26.00	45.00	52.00	53.19	63.00	129.00

Age_groups <dbl>	count <int>
20	15
30	286
40	620
50	620
60	465
70	229
80	2
120	3



What age group purchased the most fruit on average?

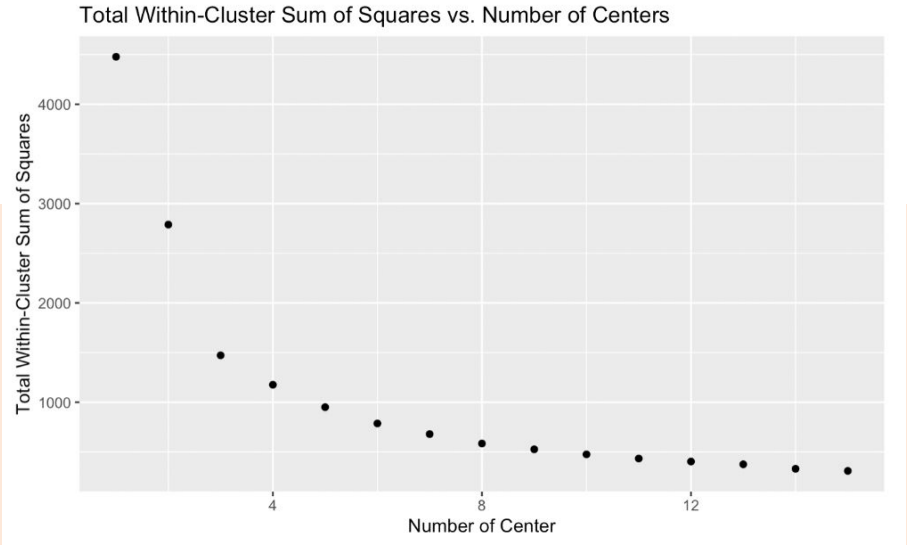


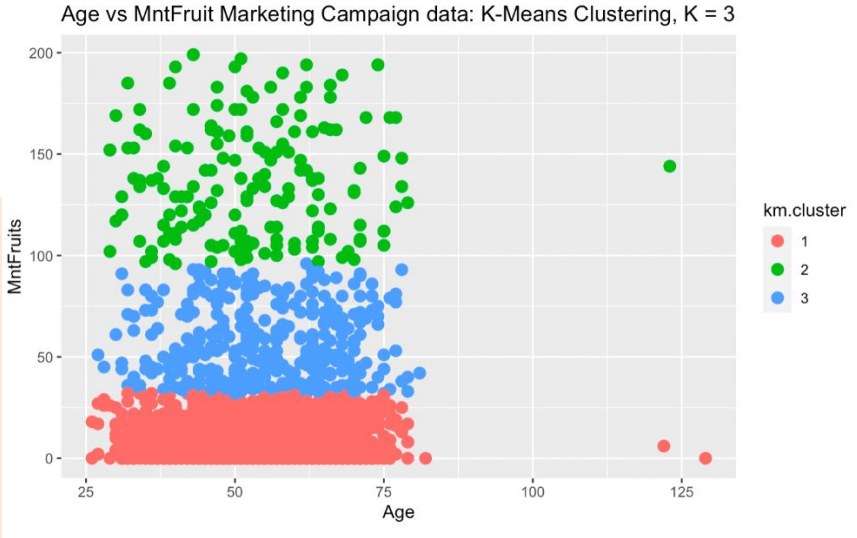


Conclusion 1: Between the 5 age groups: 70 to 79 max on average

Conclusion 2: Between middle and older age: older age max on average

What age group purchased the most fruit on average?





3 clusters are optimal

Conclusion 3: No difference among age group in clusters

What age group purchased the most fruit on average?

Conclusions:

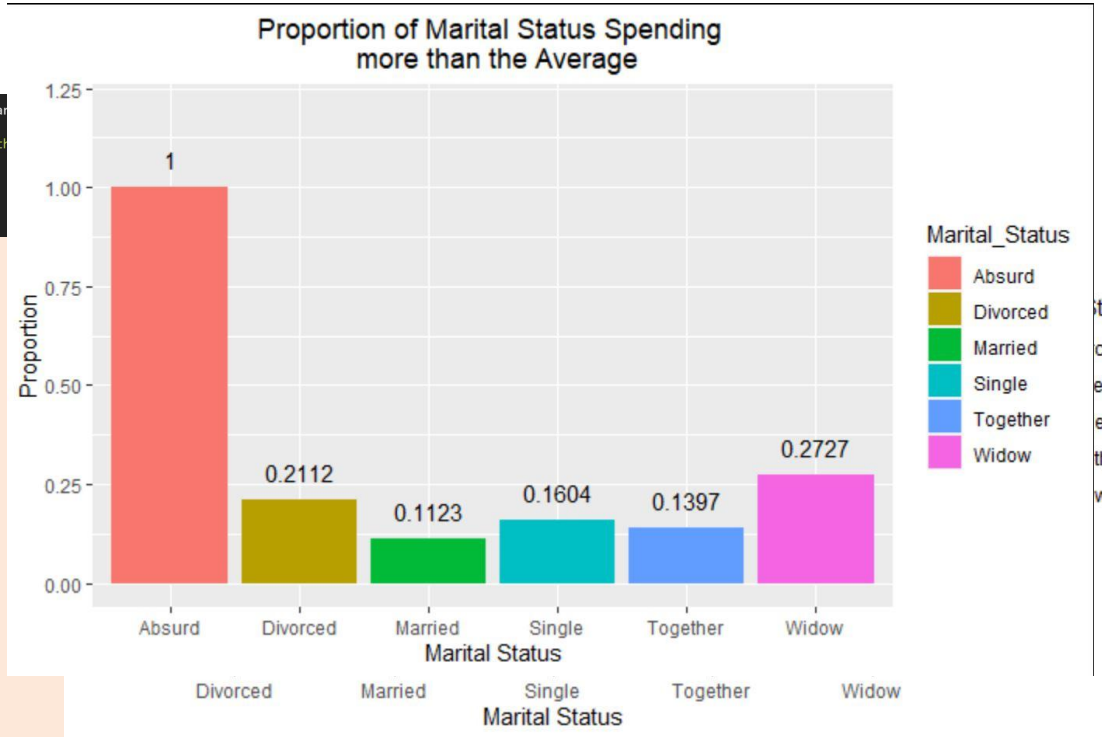
- Between the 5 age groups: 70 to 79 max on average
- Between middle and older age: older age max on average
- No difference in age between the clusters

Subquestion #3:

Are married people more likely to spend more on fruits?

Are married people more likely to spend more on fruits?

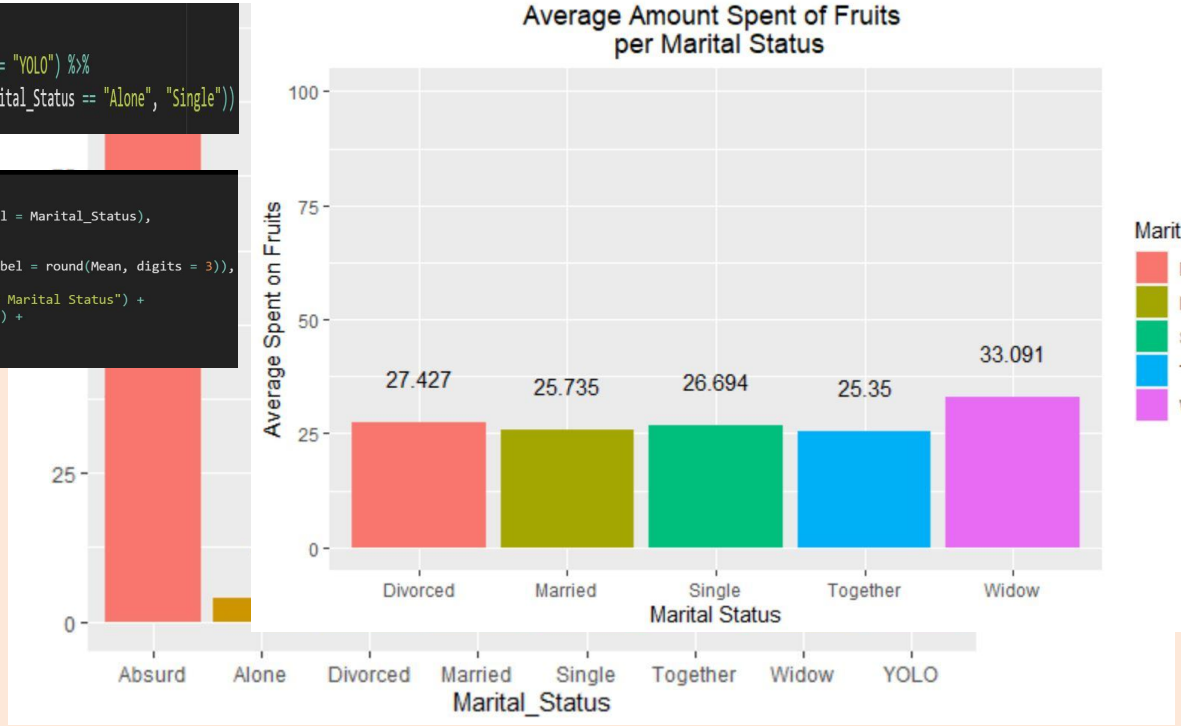
```
ggplot(rates, aes(x = Marital_Status, y = rates, fill = Mar
  geom_bar(stat = "identity") +
  ggtitle("Proportion of Marital Status Spending \n more th
    theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label = round(rates, 4)), vjust = -1) +
  ylim(0, .3) +
  xlab("Marital Status") +
  ylab("Proportion")
```



Are married people more likely to spend more on fruits?

```r
marketing_campaign.new <- marketing_campaign %>%
  filter(Marital_Status != "Absurd", Marital_Status != "YOLO") %>%
  mutate(Marital_Status = replace(Marital_Status, Marital_Status == "Alone", "Single"))
```

```r
```{r}
ggplot(new.avg.spent) +
  geom_bar(aes(x = Marital_Status, y = Mean, fill = Marital_Status),
           stat = "identity") +
  ylim(0, 100) +
  geom_text(aes(x = Marital_Status, y = Mean, label = round(Mean, digits = 3)),
            vjust = -2) +
  ggtitle("Average Amount Spent of Fruits \n per Marital Status") +
    theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Marital Status") +
  ylab("Average Spent on Fruits")
```
```



Are married people more likely to spend more on fruits?

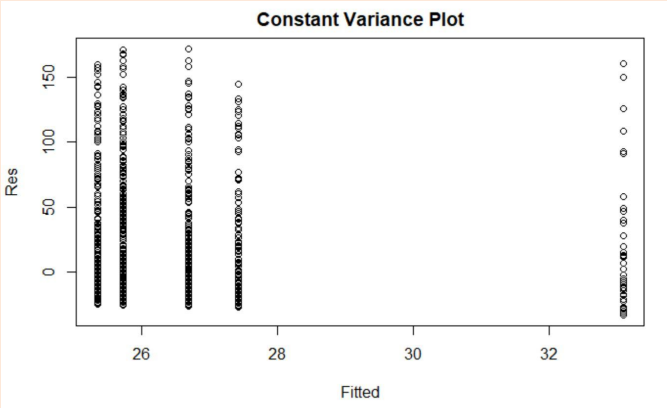
Assumptions and Results

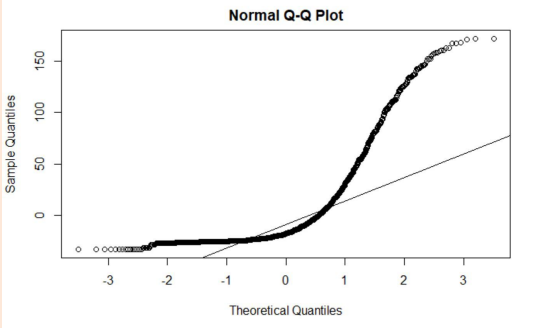
Anova to test means of Marital Status

H_0: $\mu_{Divorced} = \mu_{Married} = \mu_{Single} = \mu_{Together} = \mu_{Widowed}$

H_a: At least one μ is different

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Marital_Status	4	4718	1179	0.746	0.561
Residuals	2231	3528740	1582		





Are mar nd more on

```
        Kruskal-Wallis rank sum test

data:  MntFruits by Marital_Status
Kruskal-Wallis chi-squared = 2.5632, df = 4, p-value = 0.6333
```

Kruskal Wallis test

Comparison <chr>	Z <dbl>	P.unadj <dbl>	P.adj <dbl>
Divorced - Married	0.7294637	0.4657180	0.9981046
Divorced - Single	-0.2233368	0.8232734	1.0000000
Married - Single	-1.2634212	0.2064378	0.9009602
Divorced - Together	0.6056853	0.5447237	0.9996174
Married - Together	-0.1283316	0.8978866	1.0000000
Single - Together	1.0534267	0.2921455	0.9684180
Divorced - Widow	-0.4447660	0.6564889	0.9999771
Married - Widow	-0.9453856	0.3444620	0.9853452
Single - Widow	-0.3313160	0.7404058	0.9999986
Together - Widow	-0.8702001	0.3841910	0.9921574

Are married people more likely to spend more on fruits?

Conclusions:

- Out of the 5 groups, married people have a smaller proportion spending above the average amount for Fruits
- There is no difference between the means on ranks for the five groups.



Subquestion #4:

Is there a relationship between income and the amount spent on fruits?



Is there a relationship between income and the amount spent on fruits?

Variable “Income” had 24 missing values and 1 extreme outlier

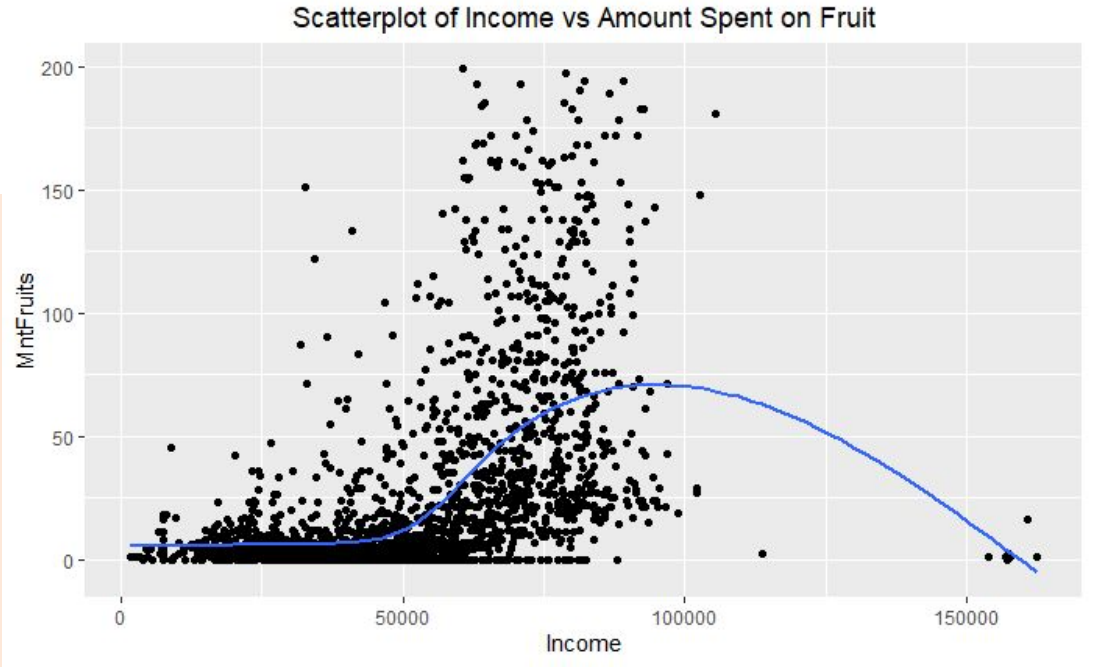
```
summary(marketing_campaign$Income)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1730   35303   51382   52247   68522  666666      24
```

Removed those rows using the dplyr function “filter”.
Extreme outlier removed for a better scatterplot and correlation
Missing values dealt with for logistic regression

Is there a relationship between income and the amount spent on fruits?

There is a correlation of 0.508 between income and among spend on fruits. There is a moderate strength linear relationship. There are a couple outliers that make the line curve down

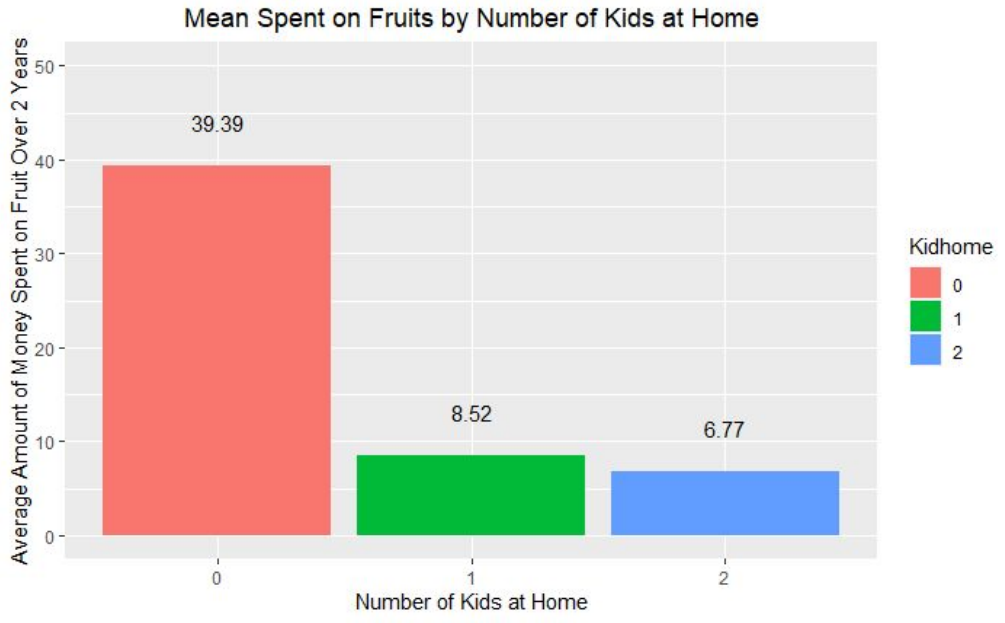


Subquestion #5:

Does the number of kids one has at home affect how much they spend on fruit?

Does the number of kids one has at home affect how much they spend on fruit?

As we can see in the below graph, customers with no kids spend around 5 times as much on fruit than customers who do have children. Those with only 1 child spend slightly more than those with 2.



Does the number of kids one has at home affect how much they spend on fruit?

The extremely small p-value from the Chi-Squared test confirms that the amount spent on fruits is dependent on the number of kids one has at home.

```
chisq.test(marketing_campaign$MntFruits, marketing_campaign$Kidhome, correct = FALSE)
```

```
        Pearson's Chi-squared test

data:  marketing_campaign$MntFruits and marketing_campaign$Kidhome
X-squared = 676.06, df = 314, p-value < 2.2e-16
```

This allows us to conclude that the amount of money spent on fruits has a dependent, negative relationship with the amount of kids a customer has at home

Main Research Question:

What characteristics makes someone more likely to purchase fruits?

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Model Assumptions:

1. Dependent variable must be binary.
2. Independent observations.
3. Independent variables should not be too highly correlated with each other. (multicollinearity)
4. No extreme outliers.
5. Linear relationship of independent variables and the logit of the dependent variable.
6. Sufficiently large sample size.

```
        Durbin-Watson test

data:  model
DW = 1.9681, p-value = 0.4536
alternative hypothesis: true autocorrelation is not 0
```

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Data Manipulation:

```
median(marketing_campaign$MntFruits)

marketing_campaign$BinaryMntFruits <- ifelse(marketing_campaign$MntFruits >= 8, 1, 0)

marketing_campaign <- marketing_campaign %>%
  mutate(Education = recode(Education,
                            "2n Cycle" = "Master's Degree",
                            "Basic" = "High School Diploma",
                            "Master" = "Master's Degree",
                            "Graduation" = "Bachelor's Degree",
                            "PhD" = "Ph.D"))

marketing_campaign <- marketing_campaign %>%
  mutate(Age = 2022 - Year_Birth)

marketing_campaign <- marketing_campaign %>%
  filter(Marital_Status != "Absurd",
         Marital_Status != "YOLO",
         Age < 90,
         Income != 666666)

marketing_campaign <- marketing_campaign %>%
  mutate(Marital_Status = ifelse(Marital_Status == "Alone", "Single", Marital_Status))

marketing_campaign$BinaryMntFruits <- as.factor(marketing_campaign$BinaryMntFruits)
marketing_campaign$Education <- as.factor(marketing_campaign$Education)
marketing_campaign$Kidhome <- as.factor(marketing_campaign$Kidhome)
```

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Full Model:

```
Call:
glm(formula = BinaryMntFruits ~ Age + Education + Marital_Status +
    Income + Kidhome, family = binomial(link = "logit"), data = marketing_campaign)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8175  -0.7560   0.2806   0.7161   2.4546

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.729e+00  3.772e-01  -4.584 4.56e-06 ***
Age                           -1.451e-02  5.014e-03  -2.893  0.00381 **
EducationHigh School Diploma   1.529e+00  3.111e-01   4.915 8.86e-07 ***
EducationMaster's Degree      -8.871e-02  1.298e-01  -0.683  0.49443
EducationPh.D                 -8.908e-01  1.393e-01  -6.394 1.62e-10 ***
Marital_StatusMarried          1.457e-02  1.874e-01   0.078  0.93800
Marital_StatusSingle           3.113e-01  2.052e-01   1.517  0.12916
Marital_StatusTogether         1.380e-01  1.966e-01   0.702  0.48258
Marital_StatusWidow            2.126e-02  3.158e-01   0.067  0.94633
Income                         6.064e-05  3.555e-06  17.058  < 2e-16 ***
Kidhome1                      -1.091e+00  1.233e-01  -8.854  < 2e-16 ***
Kidhome2                      -2.341e+00  5.475e-01  -4.275 1.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3059.5  on 2207  degrees of freedom
Residual deviance: 2168.6  on 2196  degrees of freedom
AIC: 2192.6

Number of Fisher Scoring iterations: 5
```

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Deviance Table Analysis:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: BinaryMntFruits

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          2207     3059.5
Age             1    12.70      2206     3046.8 0.0003655 ***
Education       3    19.59      2203     3027.2 0.0002064 ***
Marital_Status  4     3.22      2199     3024.0 0.5216508
Income          1   760.81      2198     2263.2 < 2.2e-16 ***
Kidhome         2    94.58      2196     2168.6 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Backwards Model Selection:

```
Start:  AIC=2192.61
BinaryMntFruits ~ Age + Education + Marital_Status + Income +
    Kidhome

                Df Deviance    AIC
- Marital_Status  4   2173.4 2189.4
<none>                2168.6 2192.6
- Age             1   2177.1 2199.1
- Education       3   2239.9 2257.9
- Kidhome         2   2263.2 2283.2
- Income          1   2548.8 2570.8

Step:  AIC=2189.4
BinaryMntFruits ~ Age + Education + Income + Kidhome

            Df Deviance    AIC
<none>           2173.4 2189.4
- Age        1   2183.3 2197.3
- Education  3   2245.3 2255.3
- Kidhome    2   2267.8 2279.8
- Income     1   2552.0 2566.0
```

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Reduced Model:

```
Call:
glm(formula = BinaryMntFruits ~ Age + Education + Income + Kidhome,
    family = binomial(link = "logit"), data = marketing_campaign)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8011  -0.7477   0.2879   0.7134   2.4500

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.560e+00  3.286e-01  -4.746 2.07e-06 ***
Age                         -1.533e-02  4.908e-03  -3.123  0.00179 **
EducationHigh School Diploma  1.549e+00  3.114e-01   4.974 6.56e-07 ***
EducationMaster's Degree    -9.426e-02  1.295e-01  -0.728  0.46662
EducationPh.D               -8.920e-01  1.392e-01  -6.408 1.48e-10 ***
Income                       6.028e-05  3.535e-06  17.053  < 2e-16 ***
Kidhome1                    -1.090e+00  1.231e-01  -8.853  < 2e-16 ***
Kidhome2                    -2.337e+00  5.484e-01  -4.262 2.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3059.5  on 2207  degrees of freedom
Residual deviance: 2173.4  on 2200  degrees of freedom
AIC: 2189.4

Number of Fisher Scoring iterations: 5
```

Estimates Interpretation:

- For every one year increase in age, the log odds of above median fruit spending decreased by 0.01532.
- Having a Highschool diploma versus Bachelor’s increased log odds of above median fruit spending by 1.549 with the same age at the same income with the same amount of kids at home respectively.

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Deviance Table Analysis:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: BinaryMntFruits

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        2207     3059.5
Age         1    12.70      2206     3046.8 0.0003655 ***
Education   3    19.59      2203     3027.2 0.0002064 ***
Income      1   759.43      2202     2267.8 < 2.2e-16 ***
Kidhome     2    94.39      2200     2173.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```




Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Odds:

(Intercept)	Age	EducationHigh School Diploma	EducationMaster's Degree
0.2102	0.9848	4.7071	0.9100
EducationPh.D	**Income**	**Kidhome1**	**Kidhome2**
0.4098	1.0001	0.3364	0.0966

- For every one year increase in age, the odds of above median fruit spending are 0.9849 fold.
- The odds of above median fruit spending with a High School diploma are 4.7 times the odds of above median fruit spending with a Bachelor's with the same gge at the same income and with the same amount of kids at home.

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Pairwise Comparison:

```
        Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: glm(formula = BinaryMntFruits ~ Age + Education + Income + Kidhome,
    family = binomial(link = "logit"), data = marketing_campaign)

Linear Hypotheses:
                                                Estimate Std. Error z value Pr(>|z|)
High School Diploma - Bachelor's Degree == 0  1.54908    0.31143   4.974  <0.0001 ***
Master's Degree - Bachelor's Degree == 0     -0.09426    0.12948  -0.728    0.877
Ph.D - Bachelor's Degree == 0                -0.89199    0.13920  -6.408  <0.0001 ***
Master's Degree - High School Diploma == 0   -1.64334    0.31963  -5.141  <0.0001 ***
Ph.D - High School Diploma == 0              -2.44107    0.33166  -7.360  <0.0001 ***
Ph.D - Master's Degree == 0                  -0.79773    0.15661  -5.094  <0.0001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

```
        Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: glm(formula = BinaryMntFruits ~ Age + Education + Income + Kidhome,
    family = binomial(link = "logit"), data = marketing_campaign)

Linear Hypotheses:
           Estimate Std. Error z value Pr(>|z|)
1 - 0 == 0  -1.0895     0.1231  -8.853   <0.001 ***
2 - 0 == 0  -2.3374     0.5484  -4.262   <0.001 ***
2 - 1 == 0  -1.2479     0.5496  -2.271   0.0508 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Cross Validation Approach:

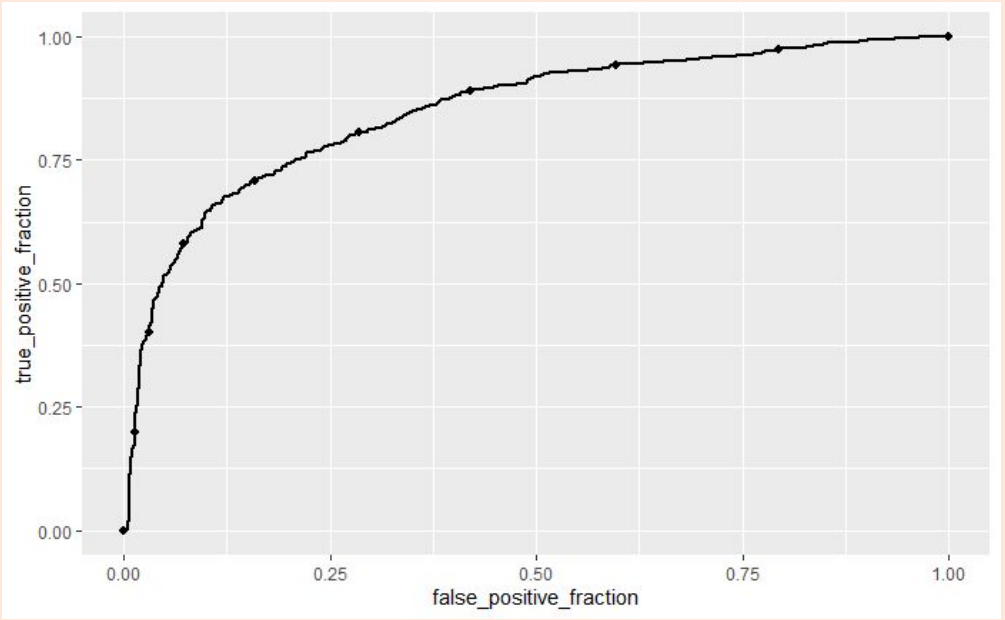
```
#cross validation approach MSE
cv.glm(marketing_campaign, reducedmodel, K = 10)$delta[1]

#weighted cross validation misclassification error rate
cost <- function(r, pi = 0) mean(abs(r-pi) > 0.5)
cv.glm(marketing_campaign, reducedmodel, cost=cost, K = 10)$delta[1]
```

- The 10-fold CV test MSE is 0.1591856.
- The weighted cross validation misclassification error rate is 0.2327899.

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

ROC Curve:



Confusion Matrix:

	class	
	0	1
0	817	259
1	253	879

Logistic Regression:
What characteristics makes someone more likely to purchase fruits?

Conclusion:

From the data analysis performed, the factors we should consider when deciding who is more likely to purchase fruits:

- Their age
- Their education level
- Their income level
- Number of kids they have at home