Thesis Data Analysis

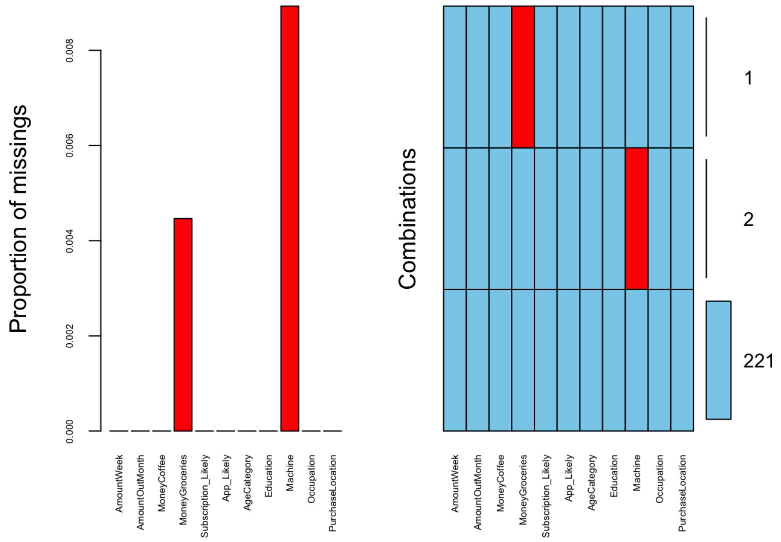
upf. Barcelona School of Management

08 March, 2021

Steps data analysis

- Univariate descriptions - categorical variables
    - Data table
    - Graphs
- Univariate descriptions - numerical variables
    - Summary
    - Confidence intervals
    - Graphs
- Boxplots - numerical
- Joint distribution tables
- Outliers
- Parametric testing
- Relationships & correlations
    - Residual plots
- Regressions
- Data problems

Introduction



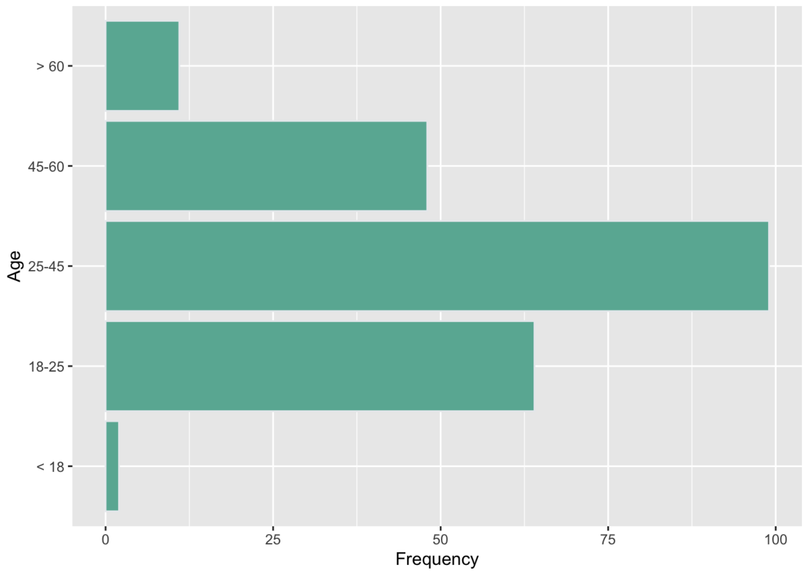
The variables included in the data set are:

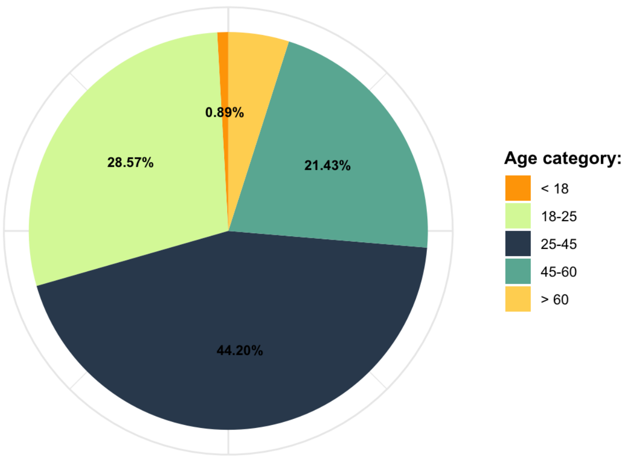
Field	Description
AmountWeek	How many cups of coffee do you typically consume weekly?
AmountOutMonth	How frequently do you drink out-of-home per month on average?
MoneyCoffee	How much money on average do you estimate you spend on coffee per month?
MoneyGroceries	How much on average do you spend on general groceries per month?
Machine	How do you brew your coffee at home?
Brand change	How often do you switch between coffee brands?
Purchase location	Where do you usually purchase your coffee?
Supermarket_Positive_Reasons	When you purchase coffee from the supermarket what are your main reasons for doing so?
Supermarket_Negative_Reasons	What would be reasons why you would not purchase coffee from the supermarket?
Criteria_Type_Coffee	What are your main criteria's or evaluation points for choosing the type of coffee?
KnowledgeCoffee	How would you describe your knowledge level regarding coffee in general?
Purchase_Price	I believe that the ____ is important to my decision on which coffee to purchase.
Purchase_Sustainability	I believe that the ____ is important to my decision on which coffee to purchase.
Purchase_Sustainability	I believe that the ____ is important to my decision on which coffee to purchase.
Purchase_Fairtrade	I believe that the ____ is important to my decision on which coffee to purchase.
Purchase_Packaging	I believe that the ____ is important to my decision on which coffee to purchase.
Frequency_Specialty	How often do you drink specialty coffee?
Subscription_Likely	How likely are you to have an online subscription for (specialty) coffee?
Subscription_Not_Likely	What is the number one reasons why you would be hesitant?
App_Likely	How likely are you to value and use an app for your online subscription?
Gender	What is your gender?
AgeCategory	What is your age category?
Occupation	What is your occupational status?
Education	What level of education have you completed?
Home	How would you describe the place you currently live in?

Univariate descriptions - Categorical variables

Age category

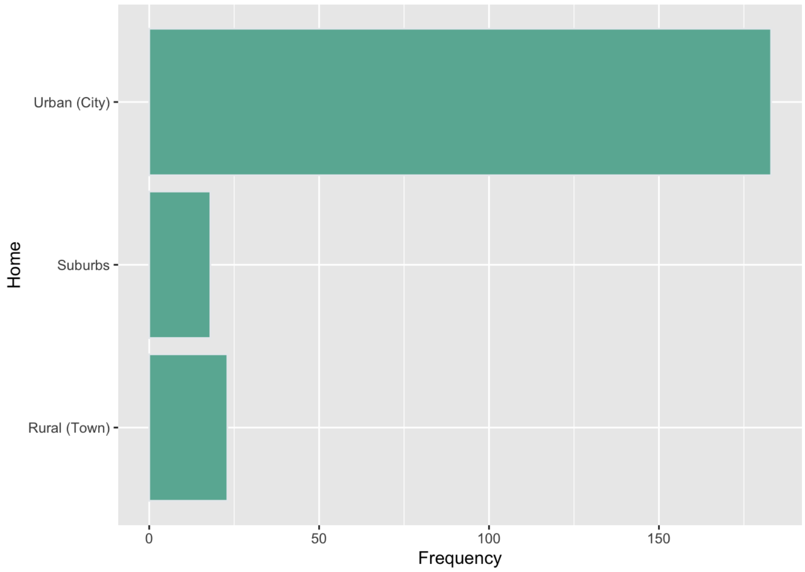
Age Category	Absolute	Relative
< 18	2	0.89%
18-25	64	28.57%
25-45	99	44.20%
45-60	48	21.43%
> 60	11	4.91%

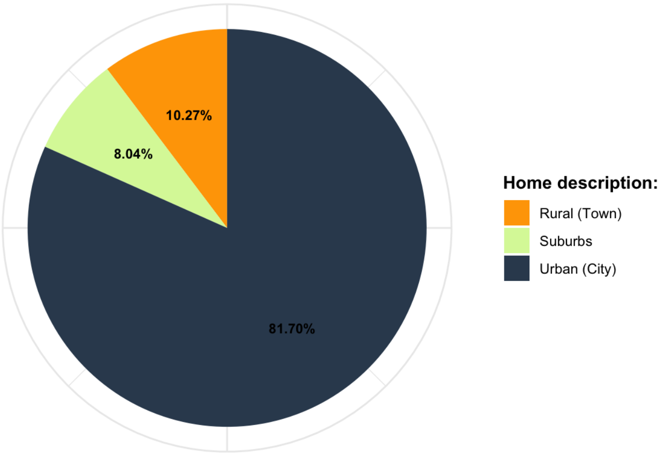




Home

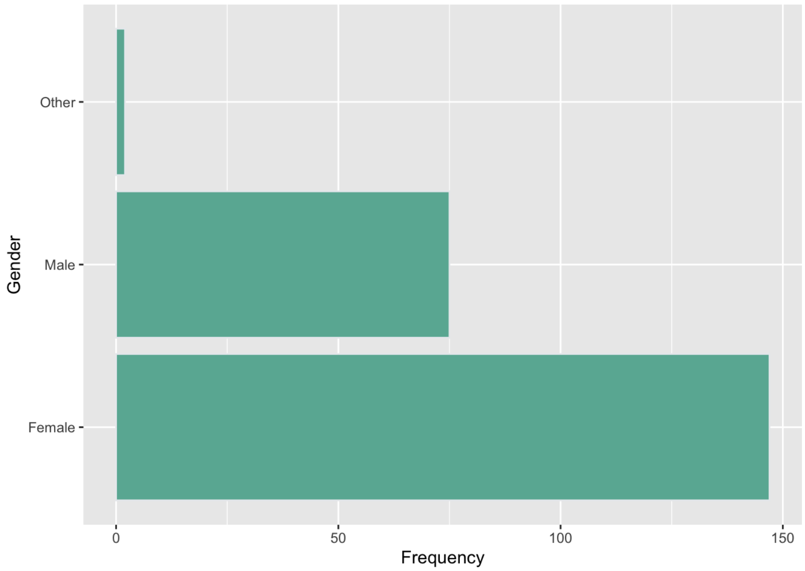
Home	Absolute	Relative
Rural (Town)	23	10.27%
Suburbs	18	8.04%
Urban (City)	183	81.70%

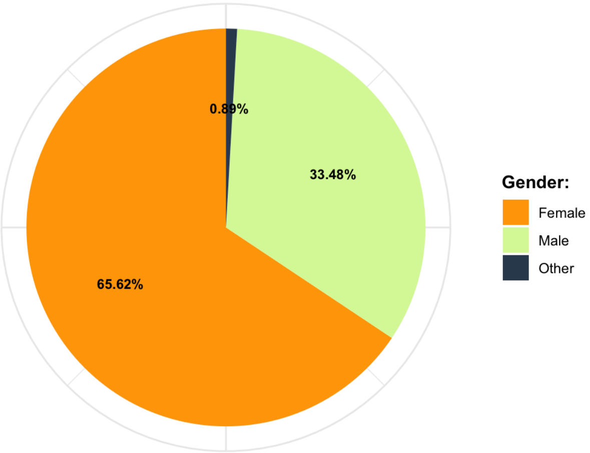




Gender

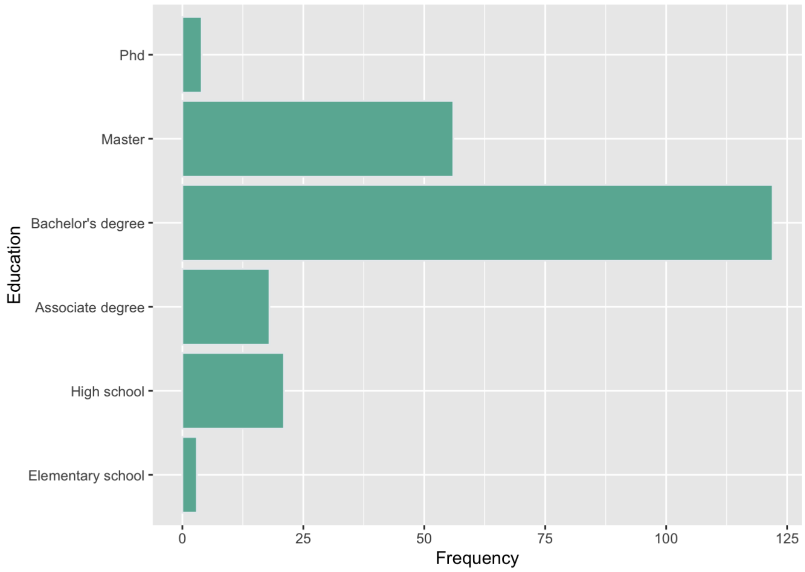
Gender	Absolute	Relative
Female	147	65.62%
Male	75	33.48%
Other	2	0.89%

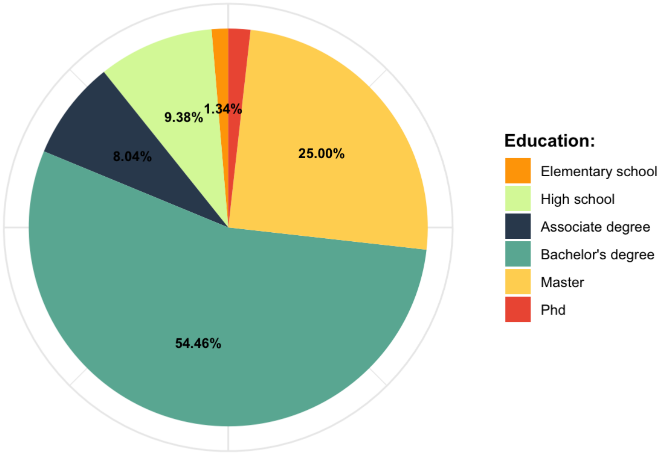




Education

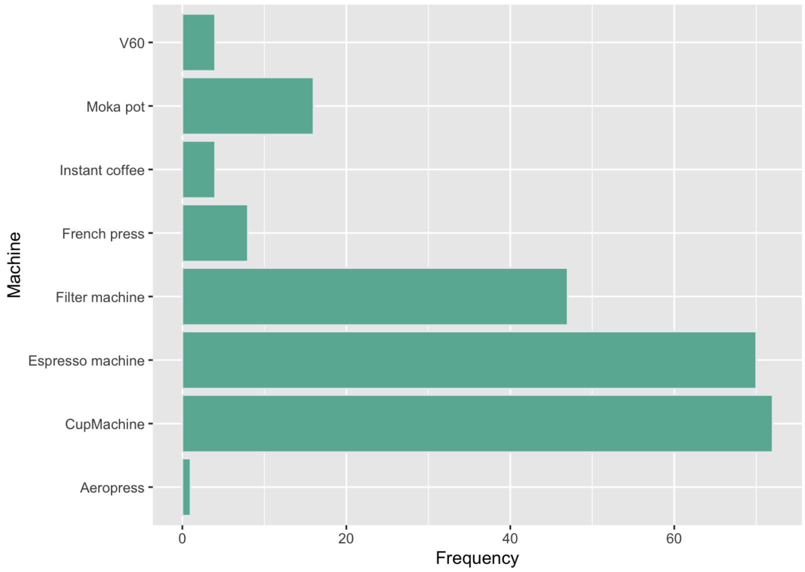
Education	Absolute	Relative
Elementary school	3	1.34%
High school	21	9.38%
Associate degree	18	8.04%
Bachelor’s degree	122	54.46%
Master	56	25.00%
Phd	4	1.79%

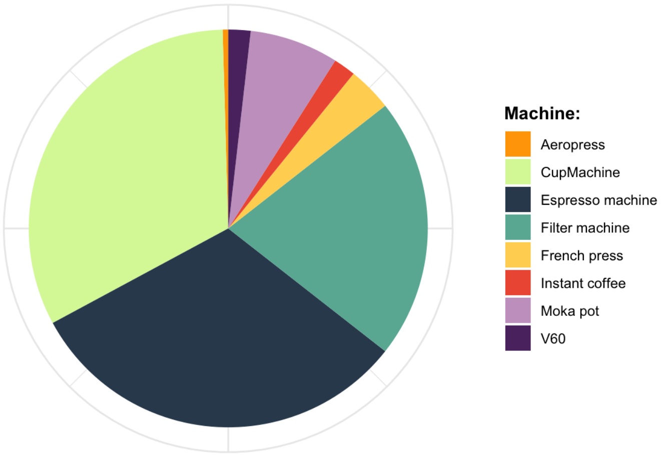




Machine

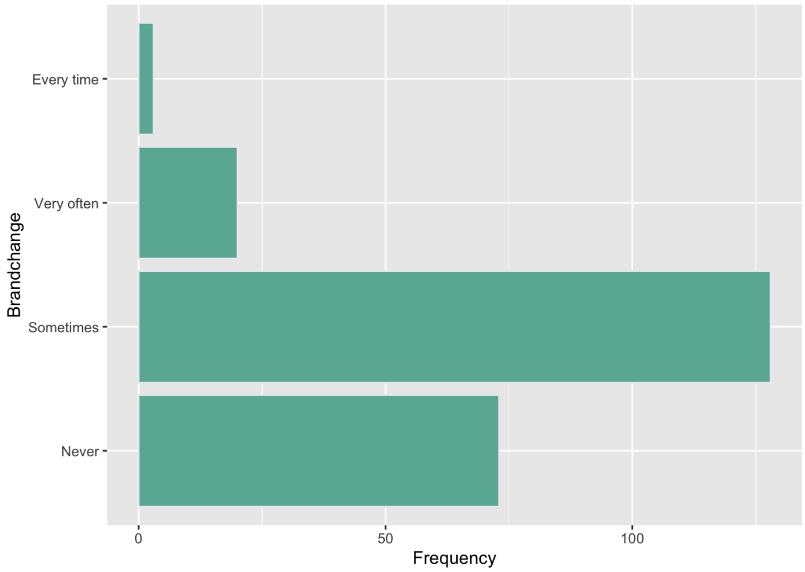
Machine	Absolute	Relative
Aeropress	1	0.45%
CupMachine	72	32.43%
Espresso machine	70	31.53%
Filter machine	47	21.17%
French press	8	3.60%
Instant coffee	4	1.80%
Moka pot	16	7.21%
V60	4	1.80%

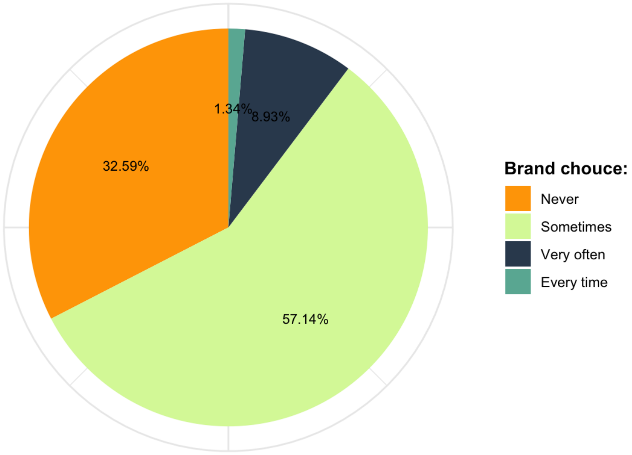




Brand choose

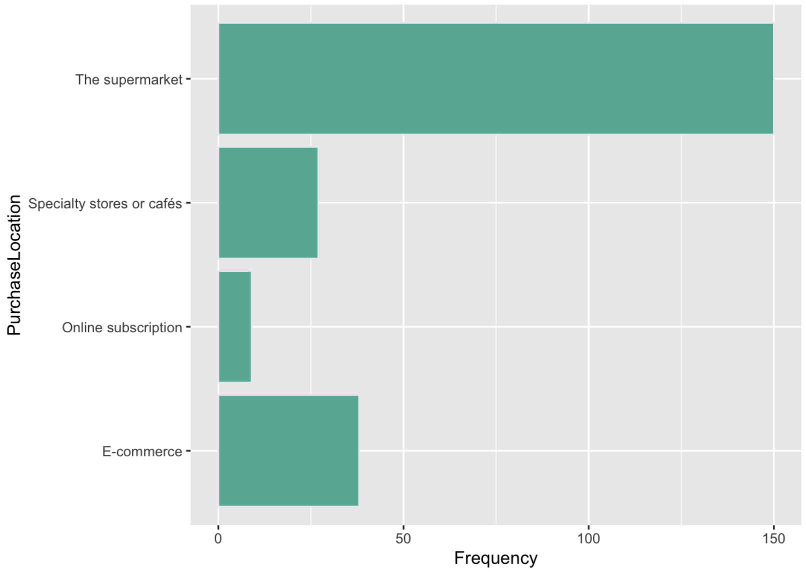
Brand choice	Absolute	Relative
Never	73	32.59%
Sometimes	128	57.14%
Very often	20	8.93%
Every time	3	1.34%

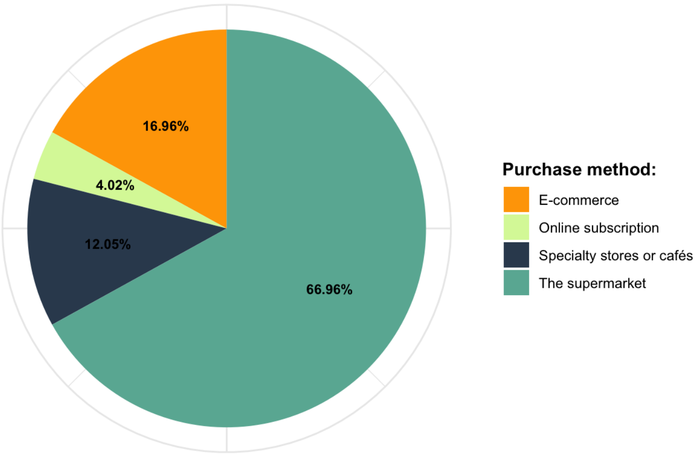




Purchase Method

Purchase Method	Absolute	Relative
E-commerce	38	16.96%
Online subscription	9	4.02%
Specialty stores or cafés	27	12.05%
The supermarket	150	66.96%





Multiple option answers:

Reasons buying from the supermarket

value	Freq
Convenience	1
I am satisfied with the product	39
I do not have special stores near where I live	8
I do not purchase coffee from the supermarket	1
Other	1
Price	14
Time-saving	32

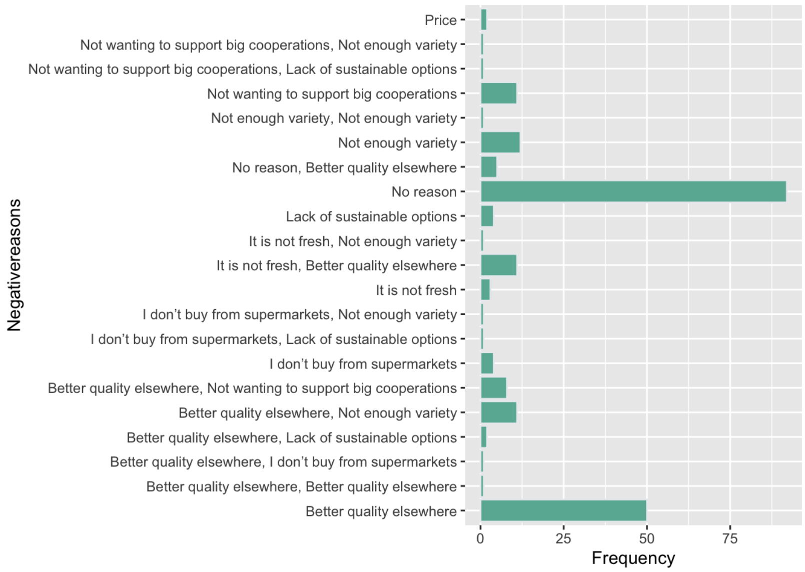
Convenience	49
I am satisfied with the product	49
I do not have special stores near where I live	8
I do not purchase coffee from the supermarket	39
Other	2
Price	55
Time-saving	22

Reasons for not buying from the supermarket

Reason	Frequency
Better quality elsewhere	17
I don’t buy from supermarkets	1
Lack of sustainable options	4
Not enough variety	15
Not wanting to support big cooperations	8
Better quality elsewhere	73
I don’t buy from supermarkets	6
It is not fresh	15
Lack of sustainable options	4
No reason	97
Not enough variety	13
Not wanting to support big cooperations	13
Price	2

Negative reasons	Absolute	Relative
Better quality elsewhere	50	22.42%
Better quality elsewhere, Better quality elsewhere	1	0.45%
Better quality elsewhere, I don’t buy from supermarkets	1	0.45%
Better quality elsewhere, Lack of sustainable options	2	0.90%
Better quality elsewhere, Not enough variety	11	4.93%
Better quality elsewhere, Not wanting to support big cooperations	8	3.59%
I don’t buy from supermarkets	4	1.79%
I don’t buy from supermarkets, Lack of sustainable options	1	0.45%
I don’t buy from supermarkets, Not enough variety	1	0.45%
It is not fresh	3	1.35%
It is not fresh, Better quality elsewhere	11	4.93%
It is not fresh, Not enough variety	1	0.45%
Lack of sustainable options	4	1.79%
No reason	92	41.26%
No reason, Better quality elsewhere	5	2.24%
Not enough variety	12	5.38%
Not enough variety, Not enough variety	1	0.45%
Not wanting to support big cooperations	11	4.93%

Not wanting to support big cooperations, Lack of sustainable options	1	0.45%
Not wanting to support big cooperations, Not enough variety	1	0.45%
Price	2	0.90%



Criteria for choosing the type of coffee

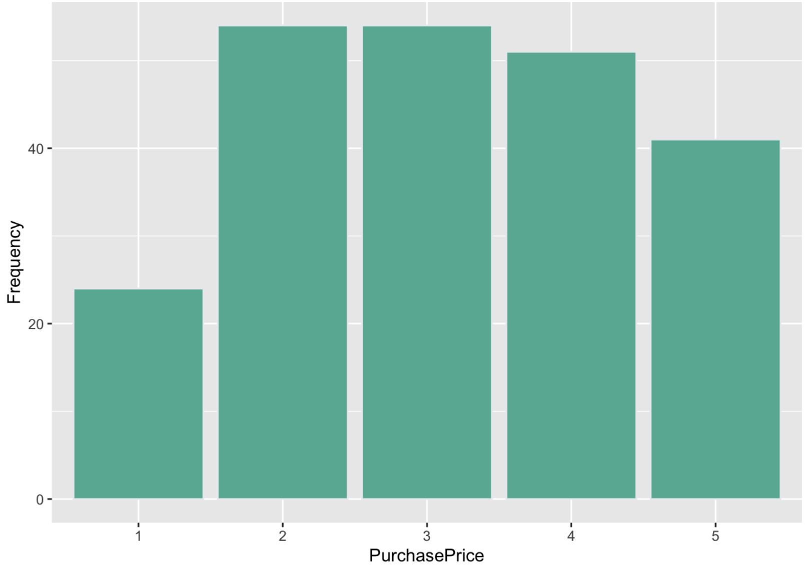
value	Freq
Sustainability & Fair Trade	3
Arabica or Robusta	8
Flavour profile	83
Origin	5
Roast level	15
Sustainability & Fair Trade	6
Arabica or Robusta	9
Flavour profile	60
Origin	31
Price	83
Roast level	46
Sustainability & Fair Trade	6

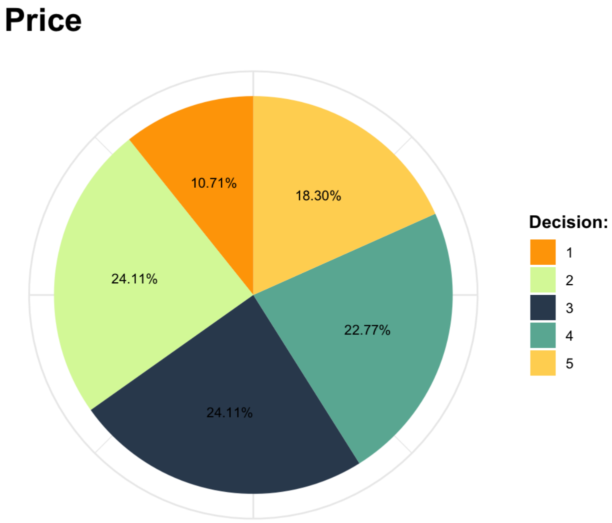
Purchase decisions 1-5

Price

Purchase decision - price	Absolute	Relative
1	24	10.71%
2	54	24.11%
3	54	24.11%
4	51	22.77%

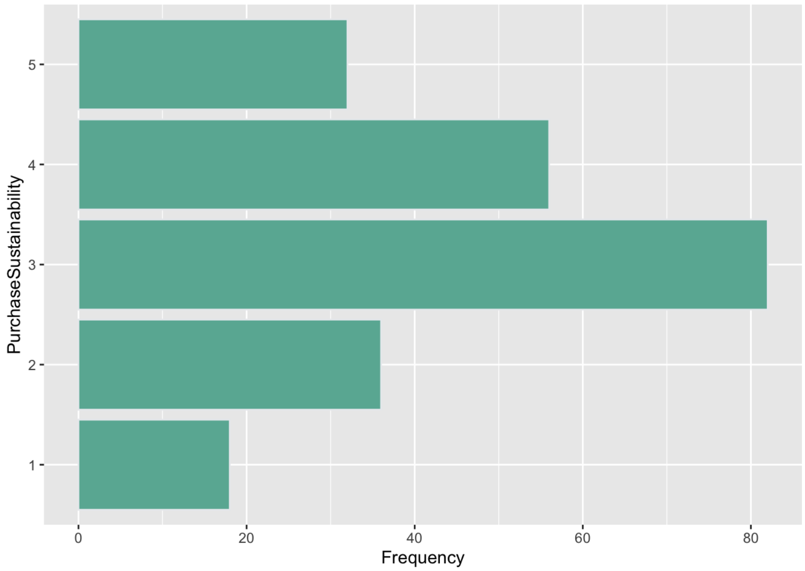
5	41	18.30%

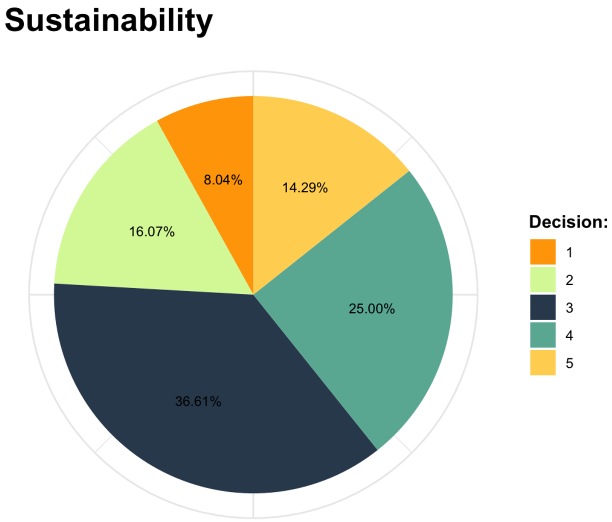




Sustainability

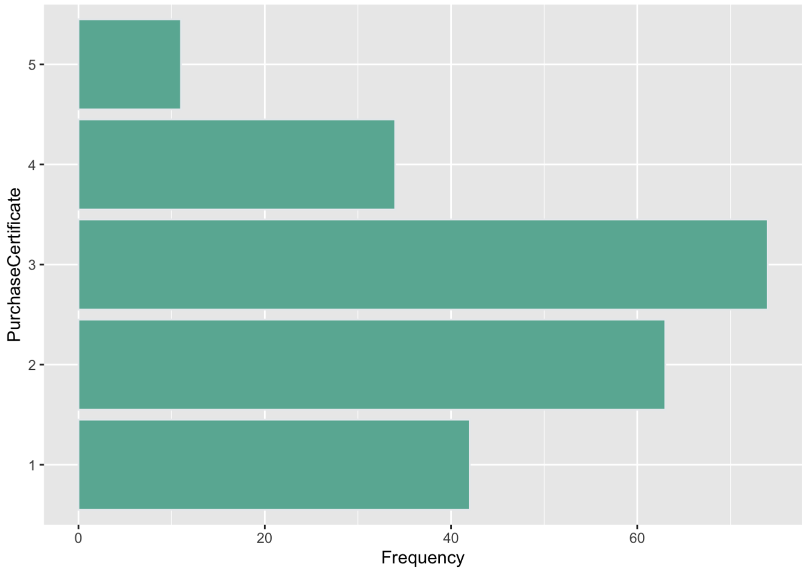
Purchase decision - sustainability	Absolute	Relative
1	18	8.04%
2	36	16.07%
3	82	36.61%
4	56	25.00%
5	32	14.29%



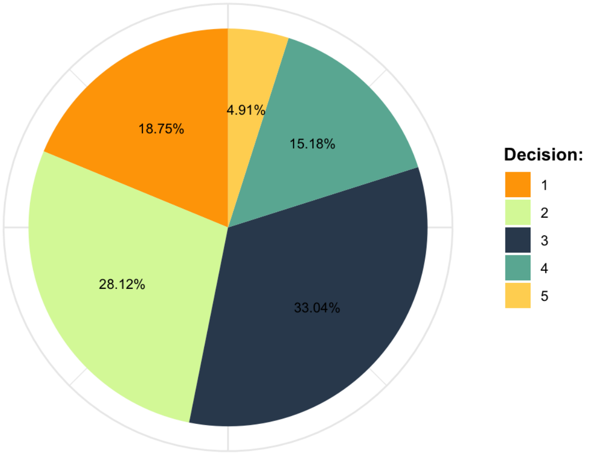

Sustainability

Certificates

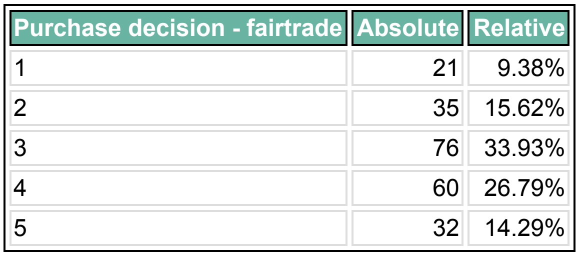
Purchase decision - certificate	Absolute	Relative
1	42	18.75%
2	63	28.12%
3	74	33.04%
4	34	15.18%
5	11	4.91%



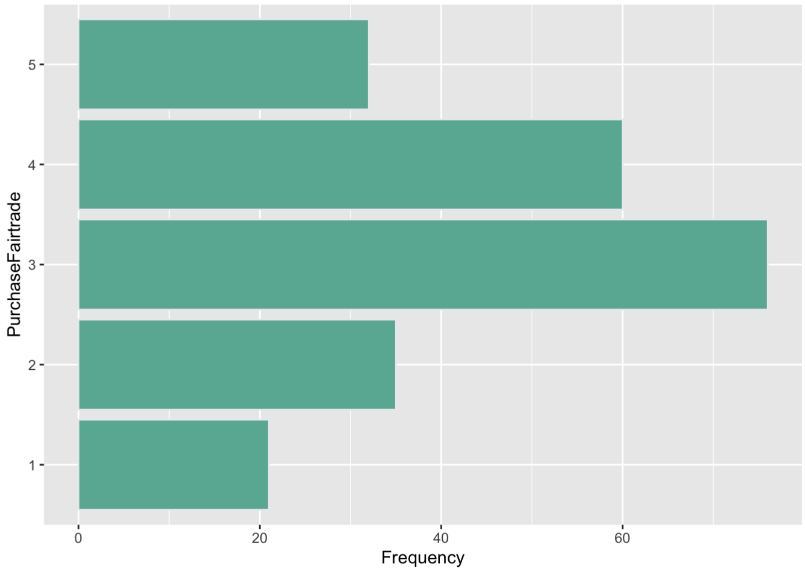
Certificate

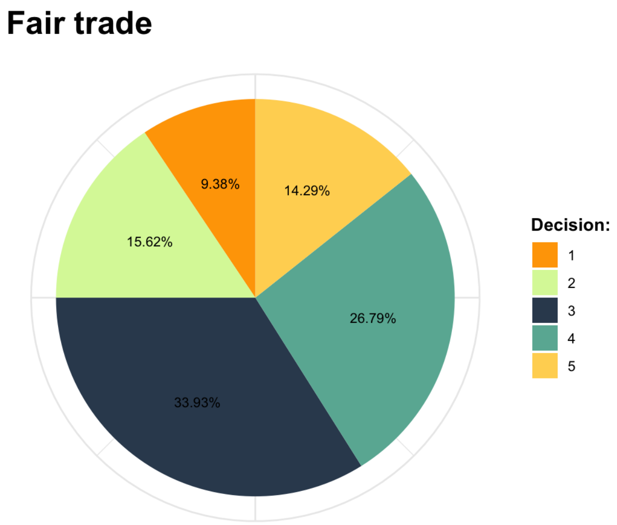


Fairtrade

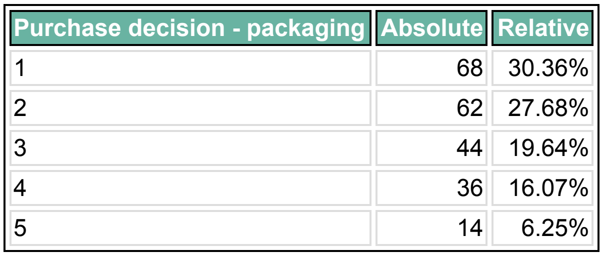


Purchase decision - fairtrade	Absolute	Relative
1	21	9.38%
2	35	15.62%
3	76	33.93%
4	60	26.79%
5	32	14.29%

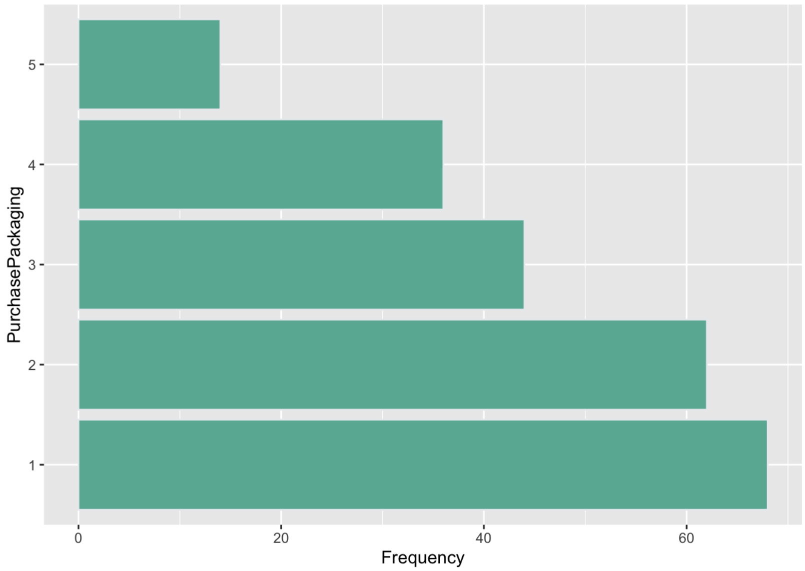




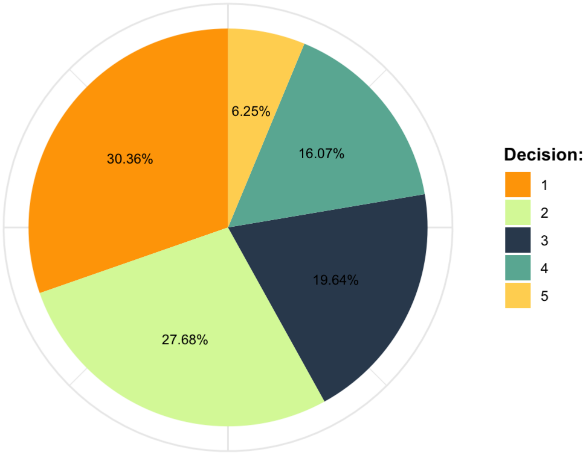
Packaging



Purchase decision - packaging	Absolute	Relative
1	68	30.36%
2	62	27.68%
3	44	19.64%
4	36	16.07%
5	14	6.25%

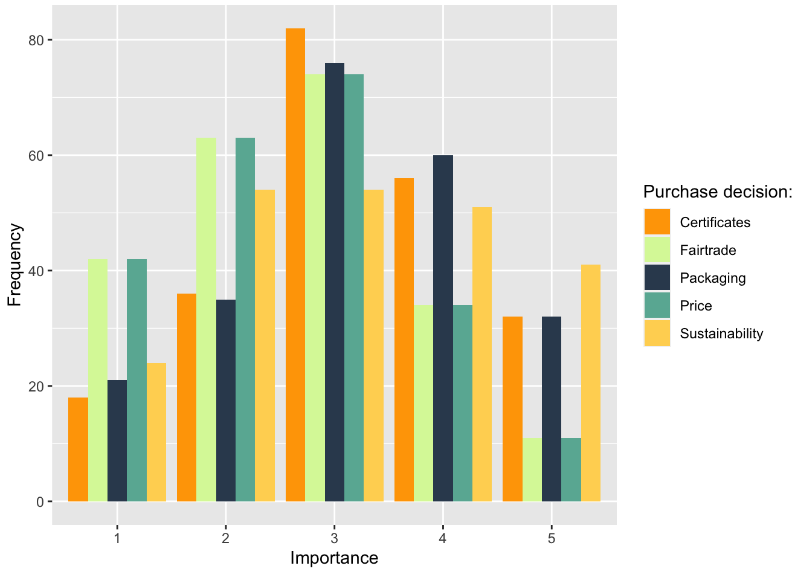


Packaging

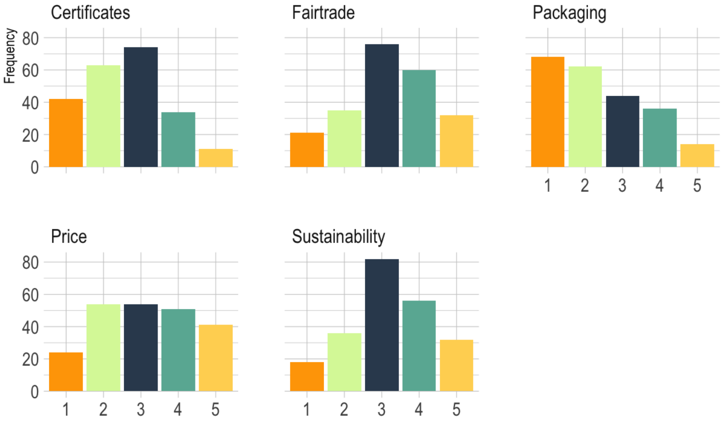


Combined data

Importance	Price	Sustainability	Certificates	Fairtrade	Packaging
1	42	24	18	42	21
2	63	54	36	63	35
3	74	54	82	74	76
4	34	51	56	34	60
5	11	41	32	11	32



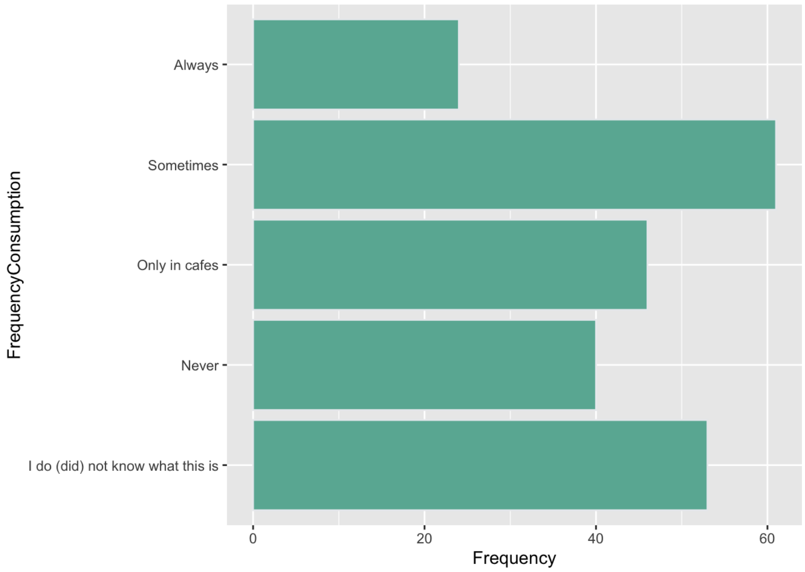
Importance of several factors on the purchasing decision.



1: Not important at all, 5: Extremely important

Frequency specialty coffee consumption

Frequency coffee consumption	Absolute	Relative
I do (did) not know what this is	53	23.66%
Never	40	17.86%
Only in cafes	46	20.54%
Sometimes	61	27.23%
Always	24	10.71%

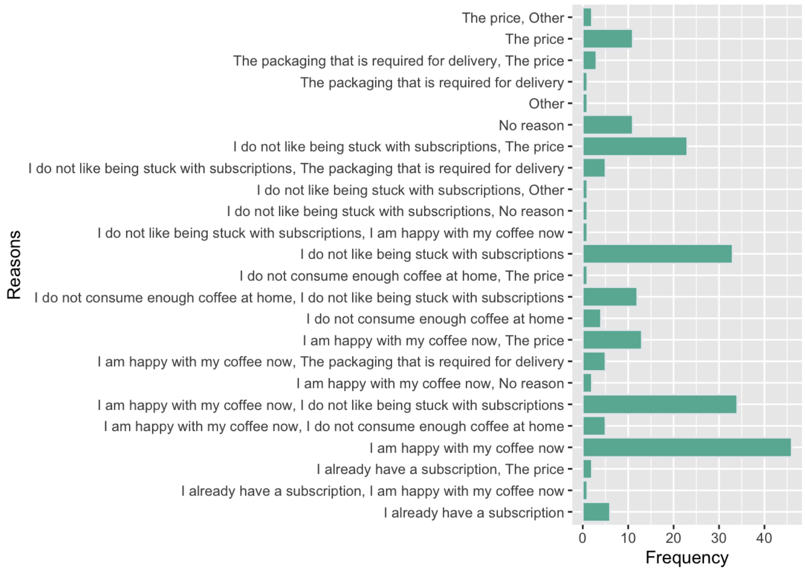


Reasons for not being likely to set up a subscription

value	Freq
I am happy with my coffee now	2
I do not consume enough coffee at home	5
I do not like being stuck with subscriptions	46
No reason	3
Other	3
The packaging that is required for delivery	10
The price	42
I already have a subscription	9
I am happy with my coffee now	105
I do not consume enough coffee at home	17
I do not like being stuck with subscriptions	64
No reason	11
Other	1
The packaging that is required for delivery	4
The price	13

Reasons	Absolute	Relative
I already have a subscription	6	2.68%
I already have a subscription, I am happy with my coffee now	1	0.45%
I already have a subscription, The price	2	0.89%
I am happy with my coffee now	46	20.54%
I am happy with my coffee now, I do not consume enough coffee at home	5	2.23%
I am happy with my coffee now, I do not like being stuck with subscriptions	34	15.18%
I am happy with my coffee now, No reason	2	0.89%
I am happy with my coffee now, The packaging that is required for delivery	5	2.23%

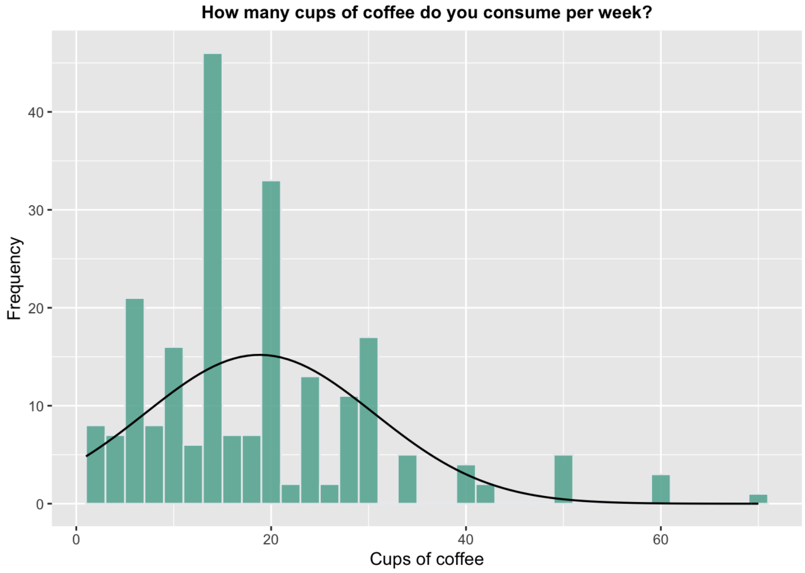
I am happy with my coffee now, The price	13	5.80%
I do not consume enough coffee at home	4	1.79%
I do not consume enough coffee at home, I do not like being stuck with subscriptions	12	5.36%
I do not consume enough coffee at home, The price	1	0.45%
I do not like being stuck with subscriptions	33	14.73%
I do not like being stuck with subscriptions, I am happy with my coffee now	1	0.45%
I do not like being stuck with subscriptions, No reason	1	0.45%
I do not like being stuck with subscriptions, Other	1	0.45%
I do not like being stuck with subscriptions, The packaging that is required for delivery	5	2.23%
I do not like being stuck with subscriptions, The price	23	10.27%
No reason	11	4.91%
Other	1	0.45%
The packaging that is required for delivery	1	0.45%
The packaging that is required for delivery, The price	3	1.34%
The price	11	4.91%
The price, Other	2	0.89%



Univariate descriptions - Numerical variables

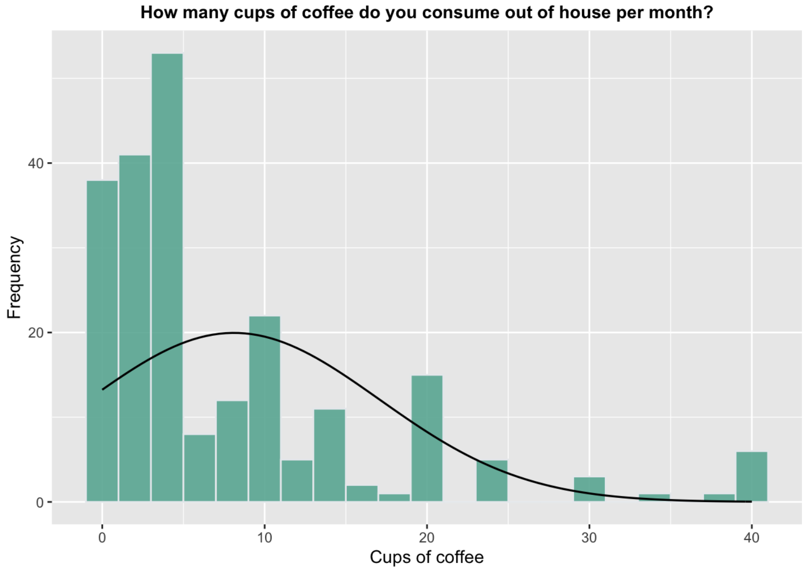
Amount coffe consumed weekly

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   10.00   15.50   18.81   25.00   70.00 
```



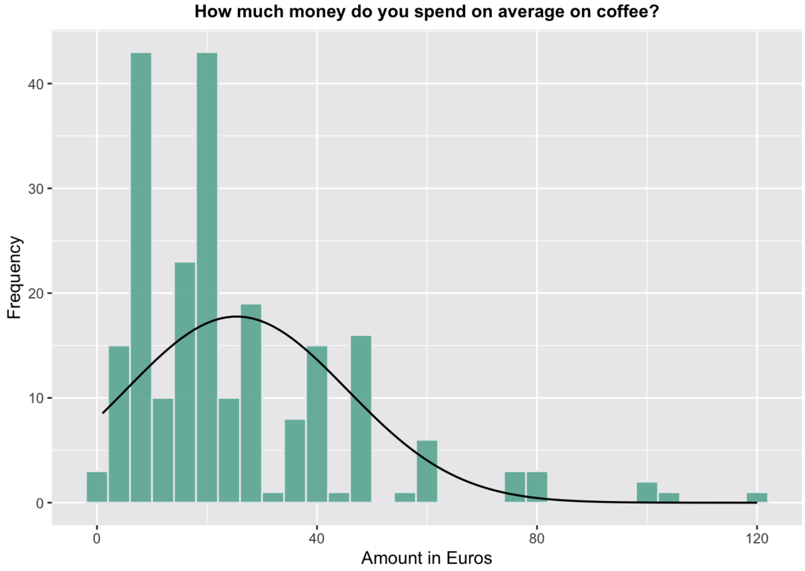
Amount per month out of house

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   2.000   5.000   8.107  10.000  40.000 
```



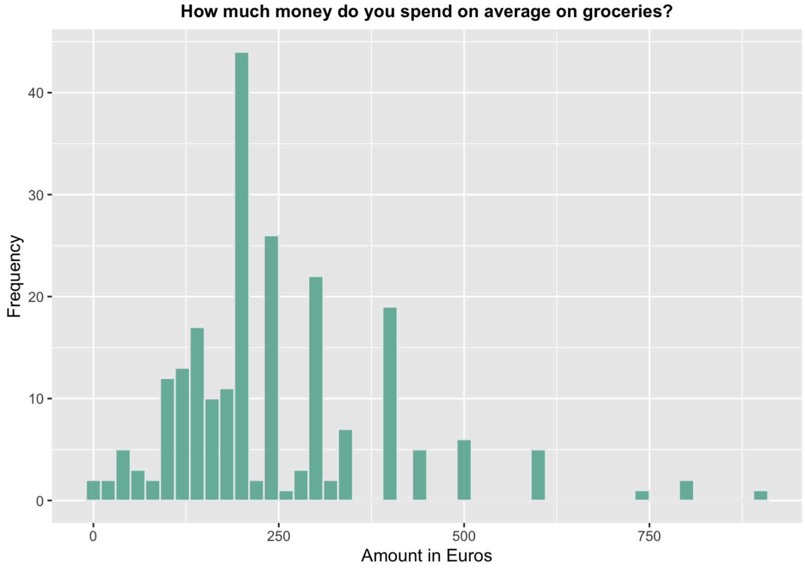
Money coffee

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   10.00   20.00   25.38   35.00  120.00 
```



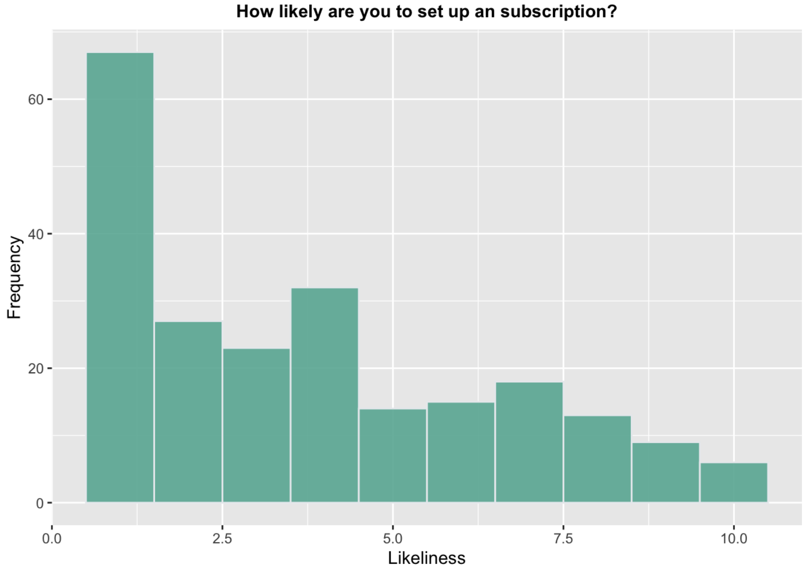
Money groceries

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    0.0   155.0   200.0   247.5   300.0   900.0       1 
```



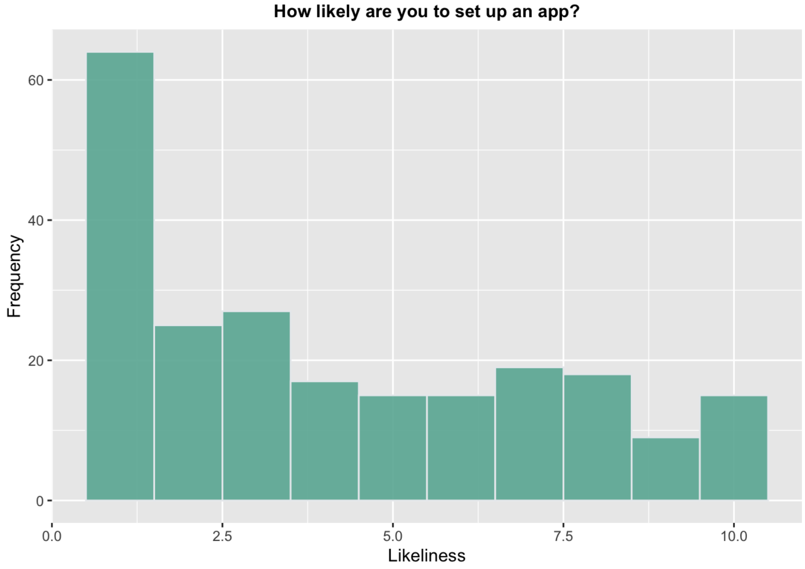
Subscription likely

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00    1.00    3.00    3.79    6.00   10.00 
```

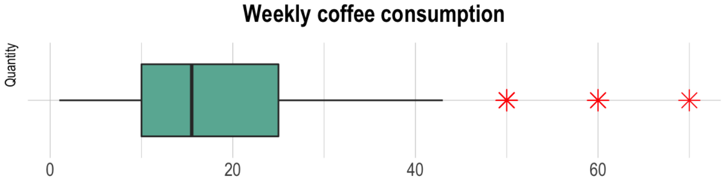


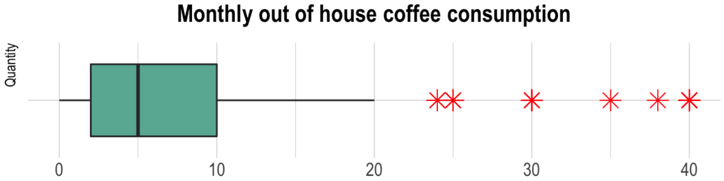
App likely

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   3.000   4.179   7.000  10.000 
```

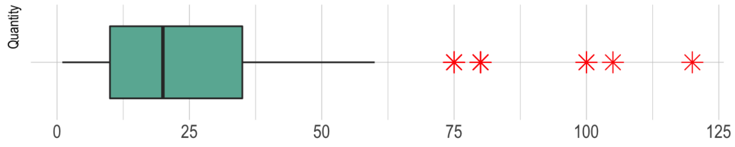


Boxplots

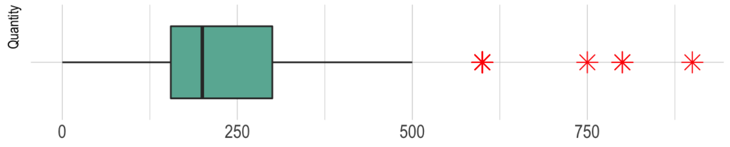

Weekly coffee consumption
Quantity
0
20
40
60


Monthly out of house coffee consumption
Quantity
0
10
20
30
40

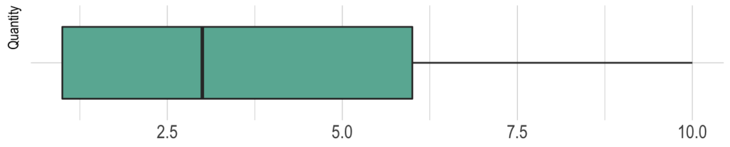
Amount of money spend on coffee



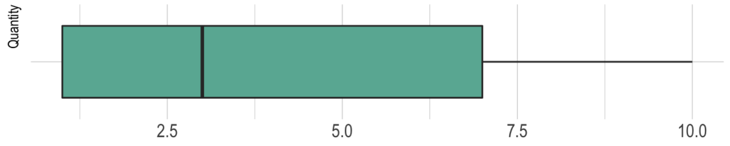
Amount spend on Groceries



How likely are you to set up an subscription



How likely are you to value and use an app?



Parametric testing

H_0 <- There is no association between the two variables.
H_a <- There is a association.

Age - Amount coffee drank

```

    Pearson's Chi-squared test

data:  AmountWeek and AgeCategory
X-squared = 254.16, df = 136, p-value = 0.000000003461


    Pearson's Chi-squared test with simulated p-value (based on 500
    replicates)

data:  AmountWeek and AgeCategory
X-squared = 254.16, df = NA, p-value = 0.00998
```

Education - Amount coffee drank

```

    Pearson's Chi-squared test

data:  AmountWeek and Education
X-squared = 236.72, df = 170, p-value = 0.000546


    Pearson's Chi-squared test with simulated p-value (based on 500
    replicates)

data:  AmountWeek and Education
X-squared = 236.72, df = NA, p-value = 0.05988
```

Gender - Amount coffee drank

```

    Pearson's Chi-squared test

data:  AmountWeek and Gender
X-squared = 71.44, df = 68, p-value = 0.3643


    Pearson's Chi-squared test with simulated p-value (based on 500
    replicates)

data:  AmountWeek and Gender
X-squared = 71.44, df = NA, p-value = 0.2954
```

Home - Amount coffee drank

```

    Pearson's Chi-squared test

data:  AmountWeek and Home
X-squared = 68.057, df = 68, p-value = 0.4753


    Pearson's Chi-squared test with simulated p-value (based on 500
    replicates)
```

```
data:  AmountWeek and Home
X-squared = 68.057, df = NA, p-value = 0.4651
```

App - Age

```

    Pearson's Chi-squared test

data:  App_Likely and AgeCategory
X-squared = 53.162, df = 36, p-value = 0.03254


    Pearson's Chi-squared test with simulated p-value (based on 500
    replicates)

data:  App_Likely and AgeCategory
X-squared = 53.162, df = NA, p-value = 0.03792
```

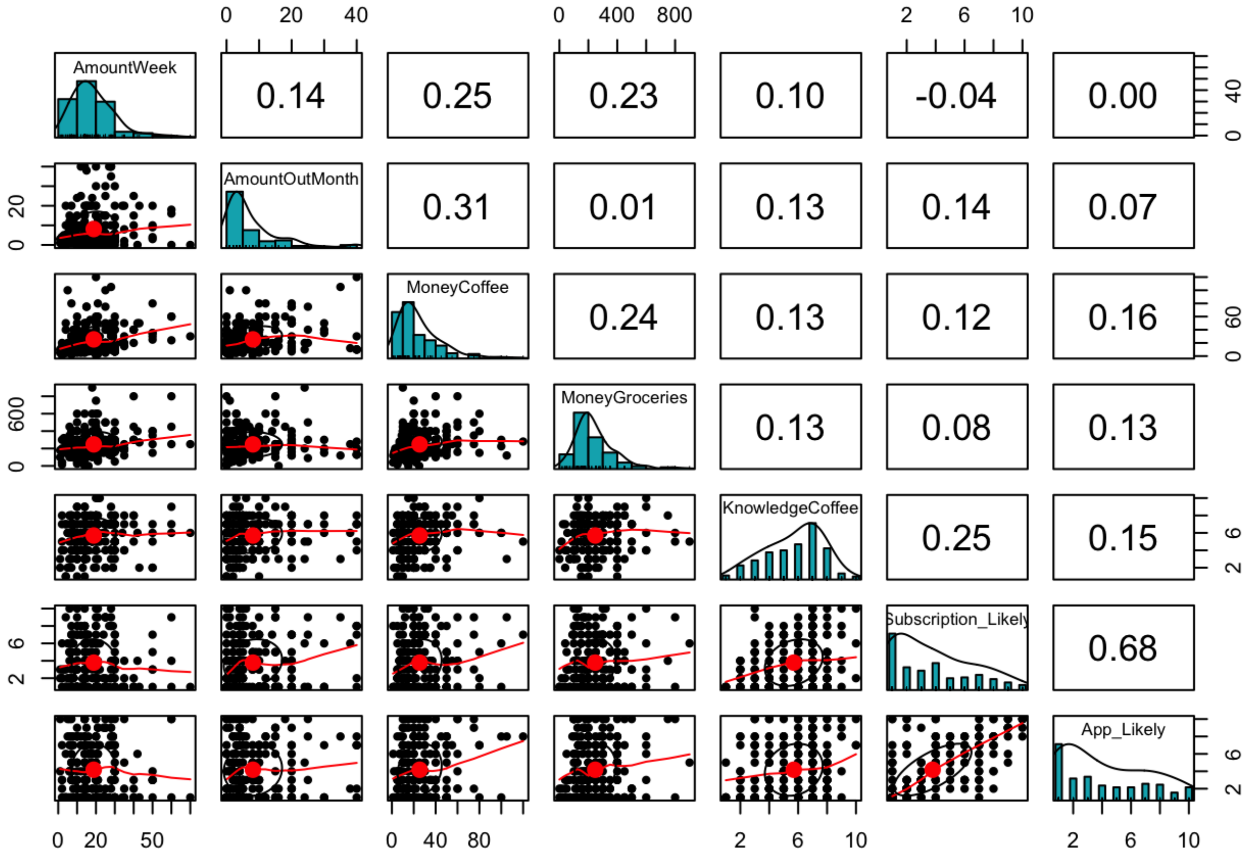
Coffee knowledge - Age

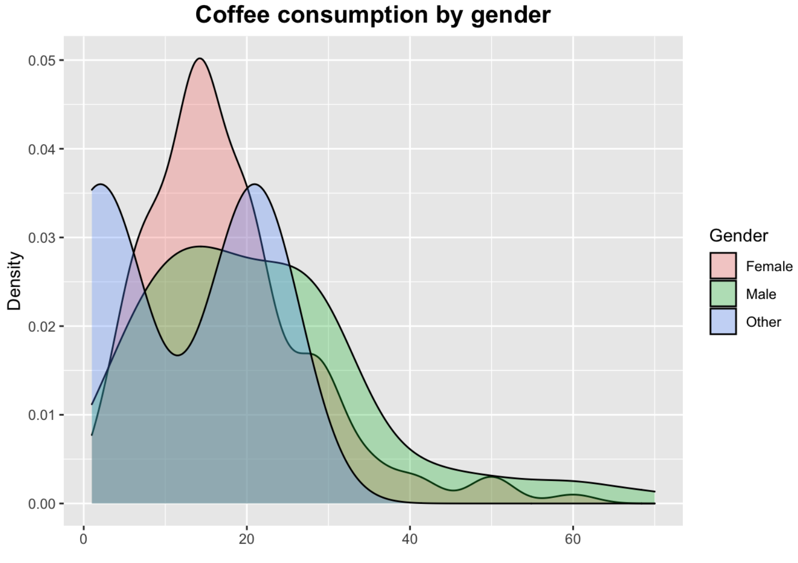
```

    Pearson's Chi-squared test

data:  KnowledgeCoffee and AgeCategory
X-squared = 151.89, df = 36, p-value = 0.0000000000000003491


    Pearson's Chi-squared test with simulated p-value (based on 500
    replicates)

data:  KnowledgeCoffee and AgeCategory
X-squared = 151.89, df = NA, p-value = 0.001996
```

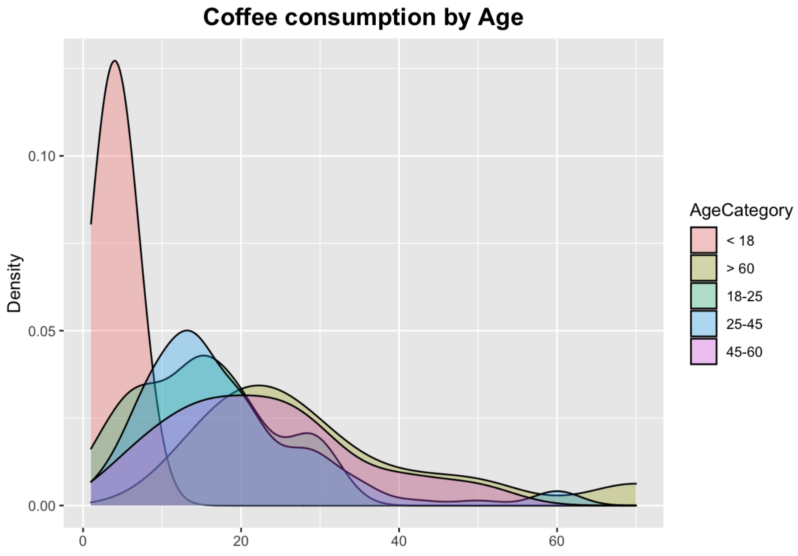
Coffee knowledge - Purchase location

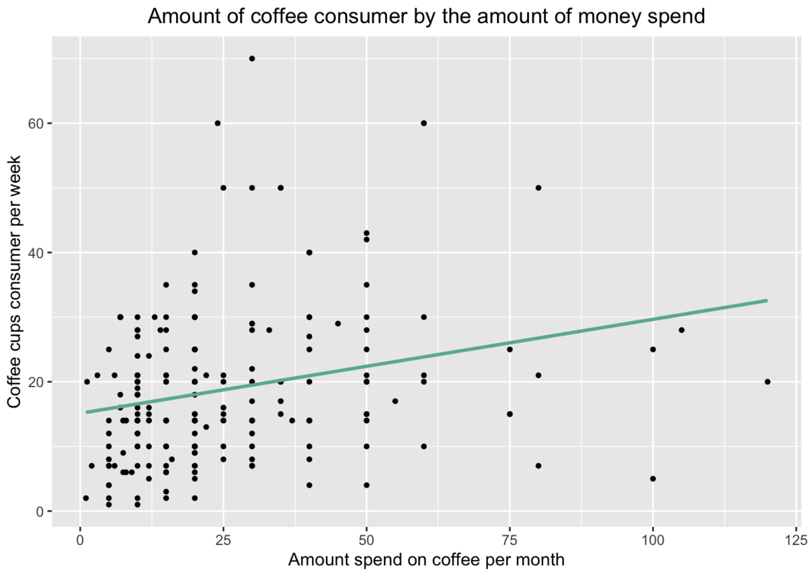
```

    Pearson's Chi-squared test

data:  KnowledgeCoffee and PurchaseLocation
X-squared = 35.066, df = 27, p-value = 0.1372


    Pearson's Chi-squared test with simulated p-value (based on 500
    replicates)

data:  KnowledgeCoffee and PurchaseLocation
X-squared = 35.066, df = NA, p-value = 0.1876
```

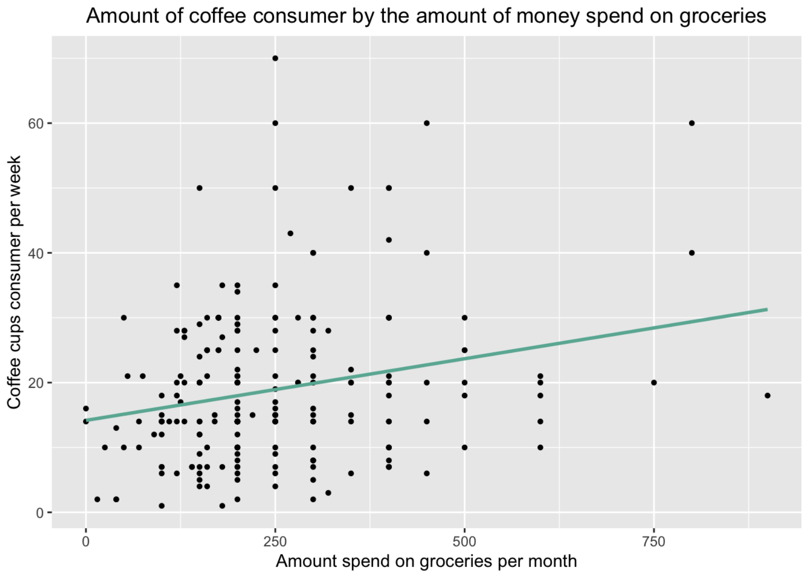
Relationships






Coffee consumption by Age
Density
0.10
0.05
0.00
0
20
40
60
AgeCategory
< 18
> 60
18-25
25-45
45-60


Amount of coffee consumer by the amount of money spend
Coffee cups consumer per week
60
40
20
0
0
25
50
75
100
125
Amount spend on coffee per month


Amount of coffee consumer by the amount of money spend on groceries
Coffee cups consumer per week
0
20
40
60
0
250
500
750
Amount spend on groceries per month

Regressions

```
Call:
lm(formula = Subscription_Likely ~ KnowledgeCoffee)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8949 -2.2229 -0.5541  2.1059  7.1115 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.88533    0.52658   3.580 0.000421 ***
KnowledgeCoffee  0.33439    0.08723   3.833 0.000165 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.607 on 222 degrees of freedom
Multiple R-squared:  0.06208,   Adjusted R-squared:  0.05785 
F-statistic: 14.69 on 1 and 222 DF,  p-value: 0.0001647
```

Incl categorical variables as dummies

Cooks distance –> outliers

Data problems