Thesis Data Analysis

upf. Barcelona School of Management

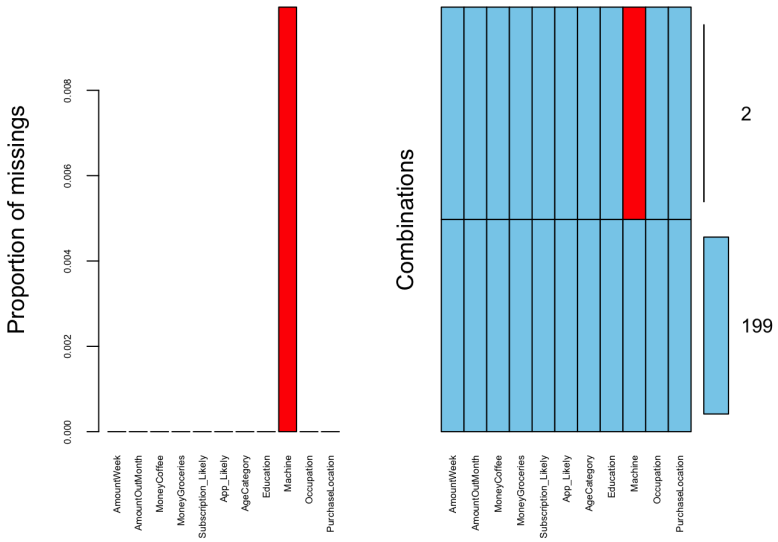
04 March, 2021

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



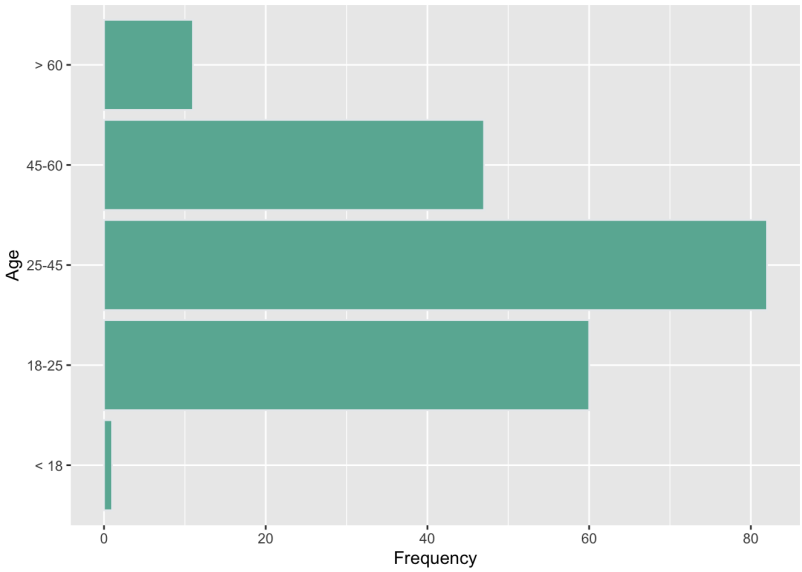
The variables included in the data set are:

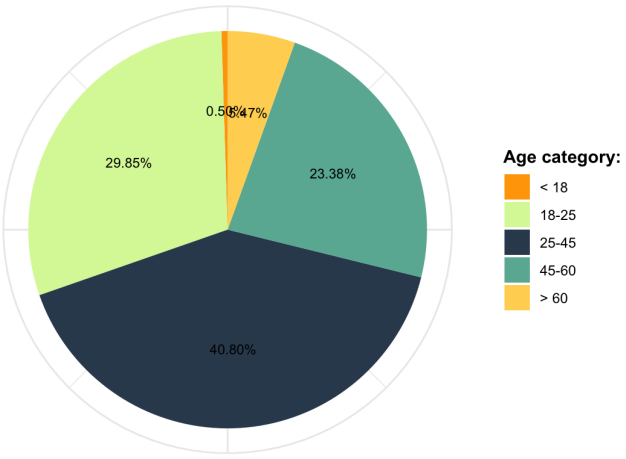
Field	Description
AmountWeek	How many cups of coffee do you typically consume weekly?
AmountOutMonth	How frequently do you drink out-of-home per month on average?
MoneyCoffee	How much money on average do you estimate you spend on coffee per month?
MoneyGroceries	How much on average do you spend on general 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

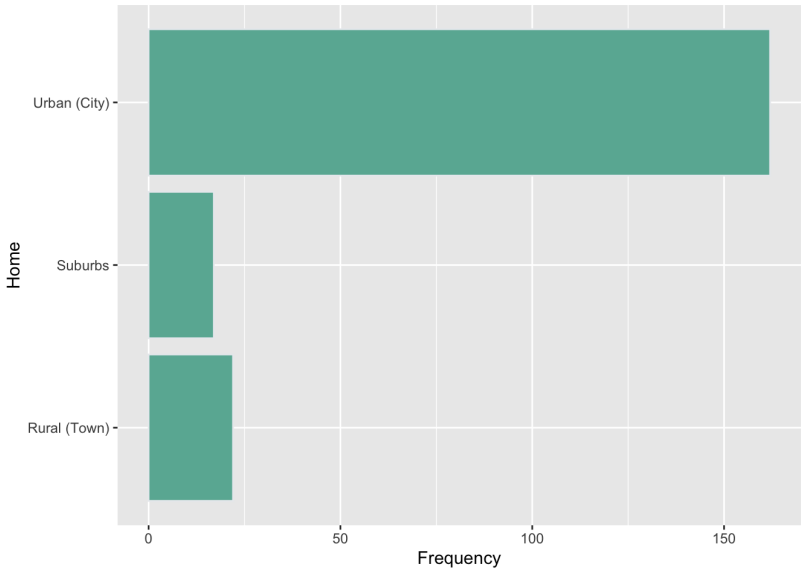
Age Category	Absolute	Relative
< 18	1	0.50%
18-25	60	29.85%
25-45	82	40.80%
45-60	47	23.38%
> 60	11	5.47%

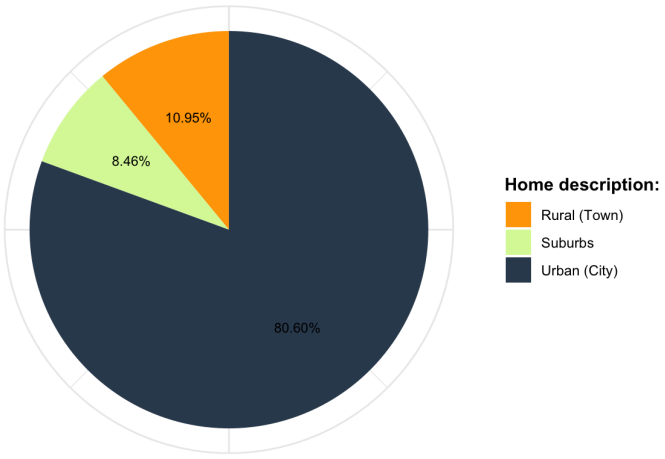




Home

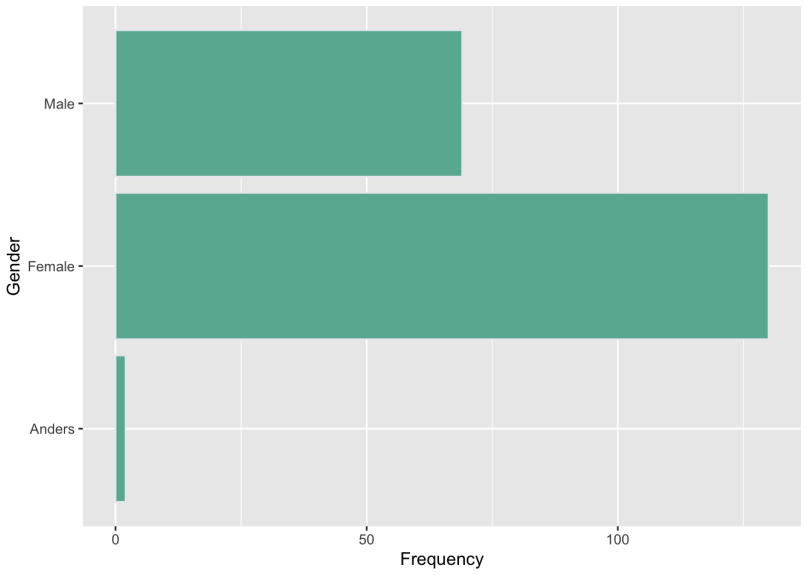
Home	Absolute	Relative
Rural (Town)	22	10.95%
Suburbs	17	8.46%
Urban (City)	162	80.60%

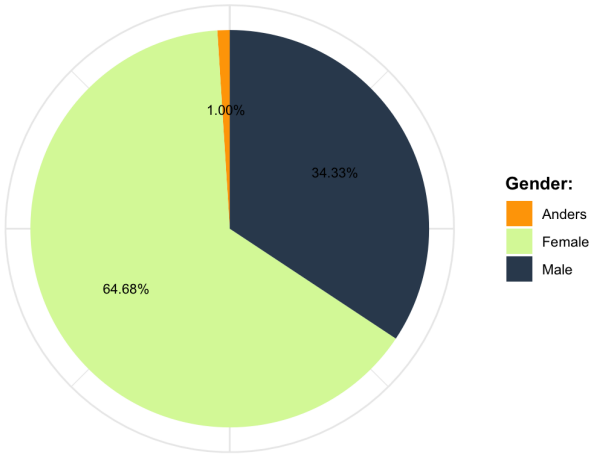




Gender

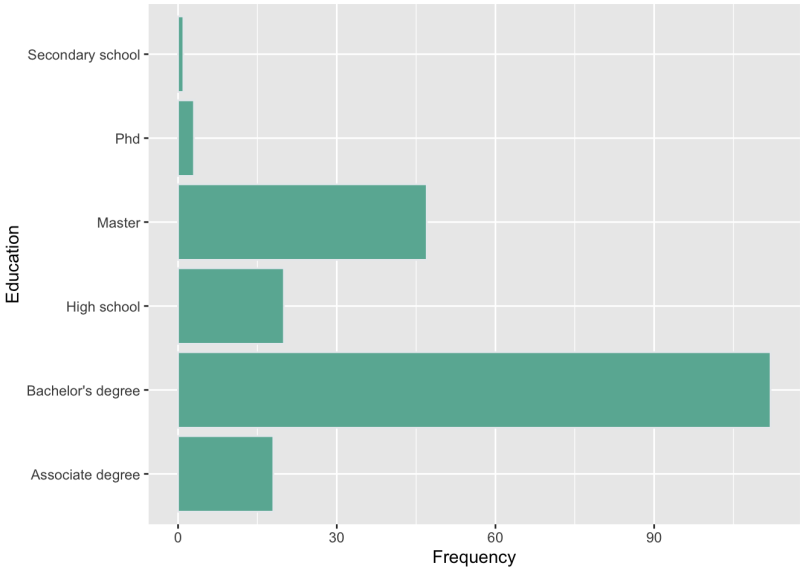
Gender	Absolute	Relative
Anders	2	1.00%
Female	130	64.68%
Male	69	34.33%

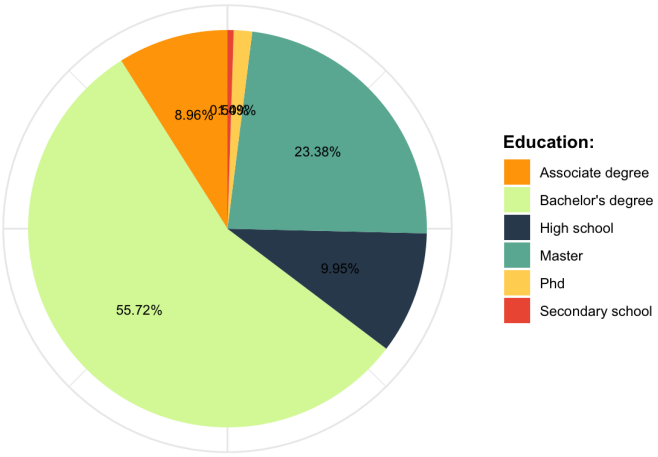




Education

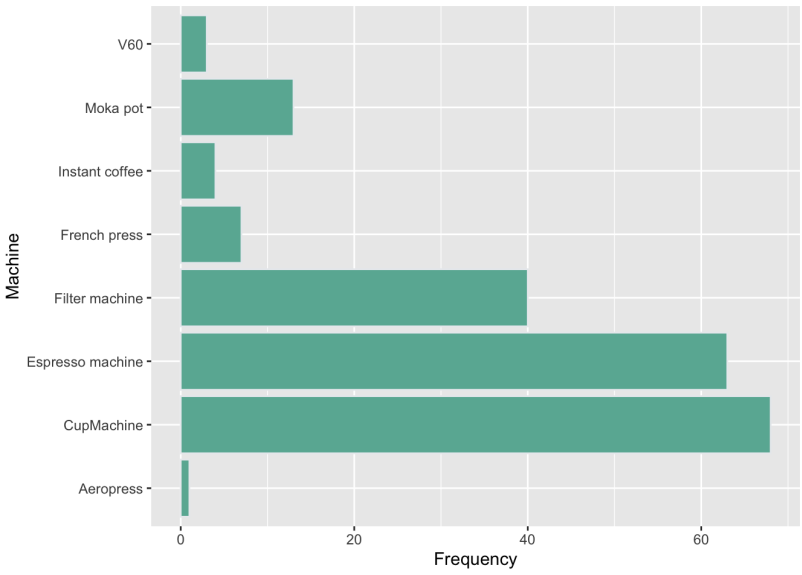
Education	Absolute	Relative
Associate degree	18	8.96%
Bachelor’s degree	112	55.72%
High school	20	9.95%
Master	47	23.38%
Phd	3	1.49%
Secondary school	1	0.50%

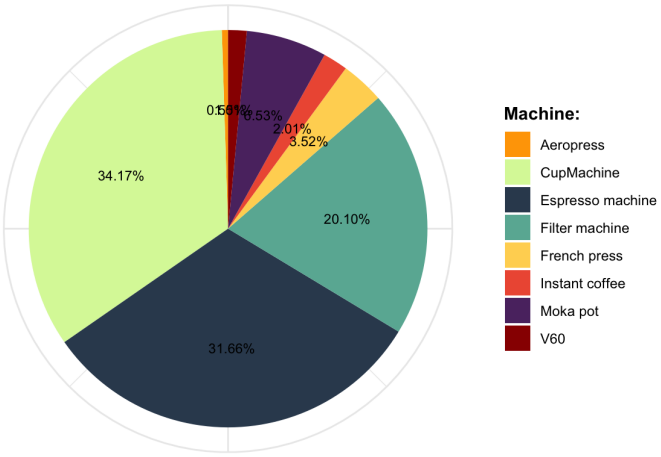




Machine

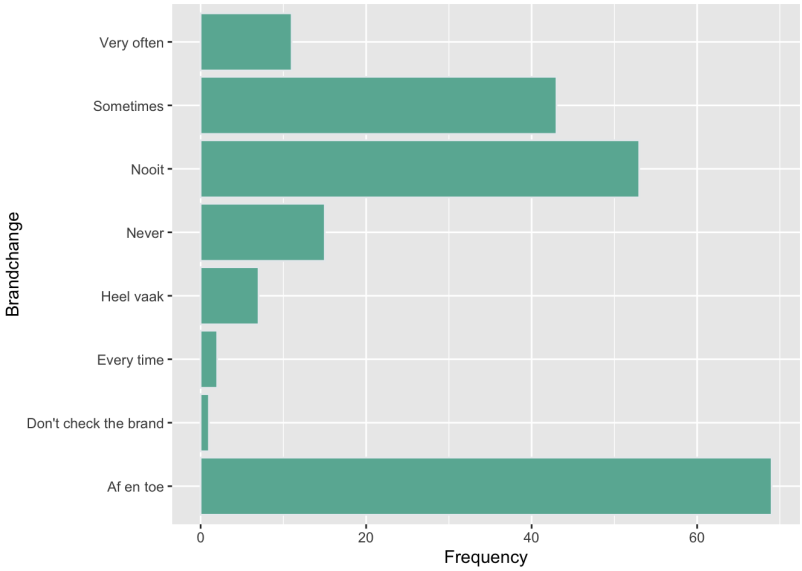
Machine	Absolute	Relative
Aeropress	1	0.50%
CupMachine	68	34.17%
Espresso machine	63	31.66%
Filter machine	40	20.10%
French press	7	3.52%
Instant coffee	4	2.01%
Moka pot	13	6.53%
V60	3	1.51%

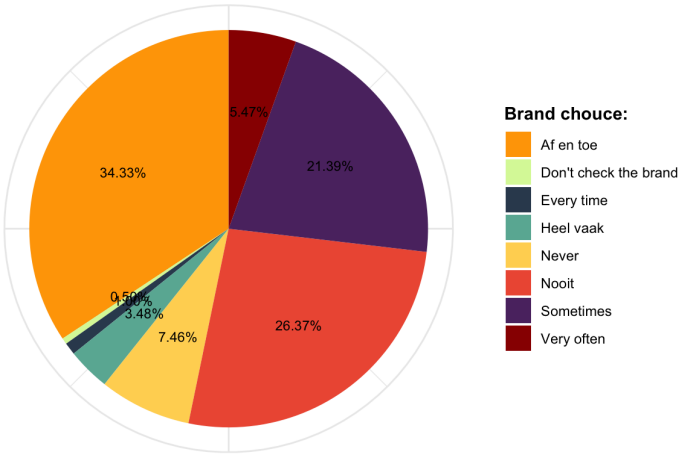




Brand choose

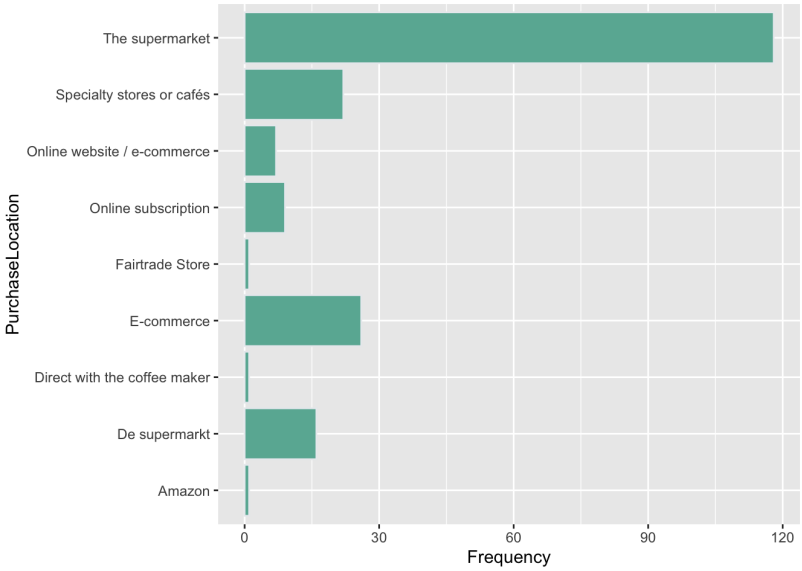
Brand choice	Absolute	Relative
Af en toe	69	34.33%
Don’t check the brand	1	0.50%
Every time	2	1.00%
Heel vaak	7	3.48%
Never	15	7.46%
Nooit	53	26.37%
Sometimes	43	21.39%
Very often	11	5.47%

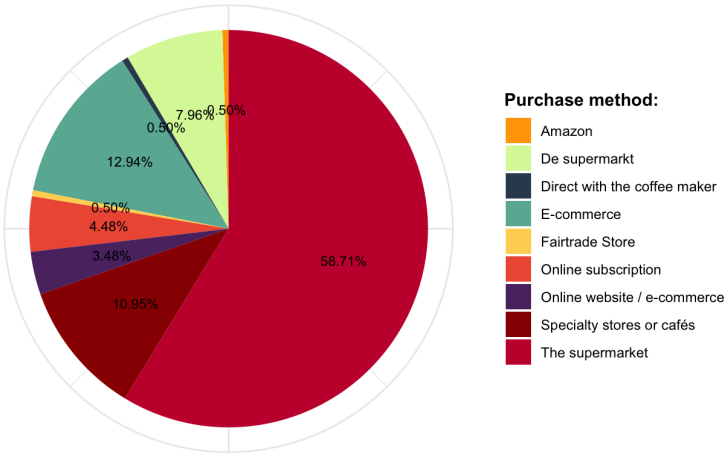




Purchase Method

Purchase Method	Absolute	Relative
Amazon	1	0.50%
De supermarkt	16	7.96%
Direct with the coffee maker	1	0.50%
E-commerce	26	12.94%
Fairtrade Store	1	0.50%
Online subscription	9	4.48%
Online website / e-commerce	7	3.48%
Specialty stores or cafés	22	10.95%
The supermarket	118	58.71%

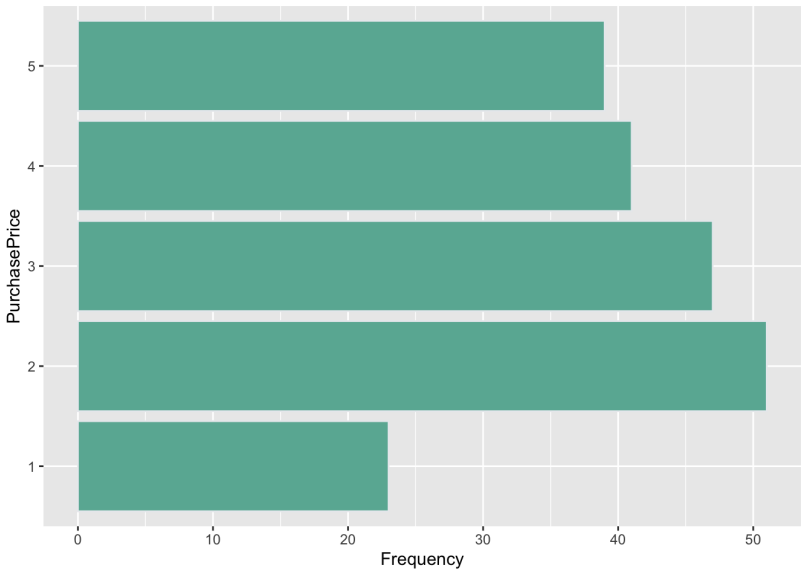


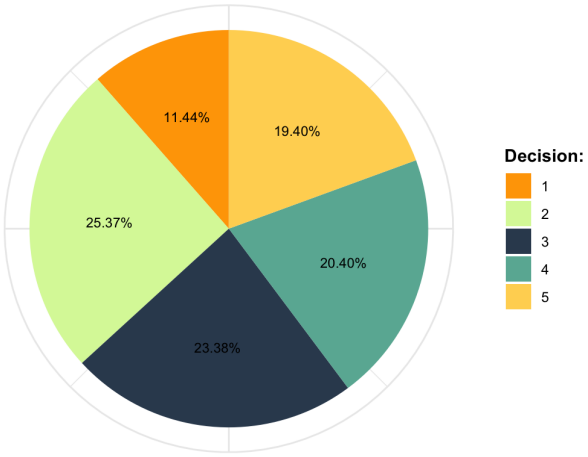


Purchase decisions 1-5

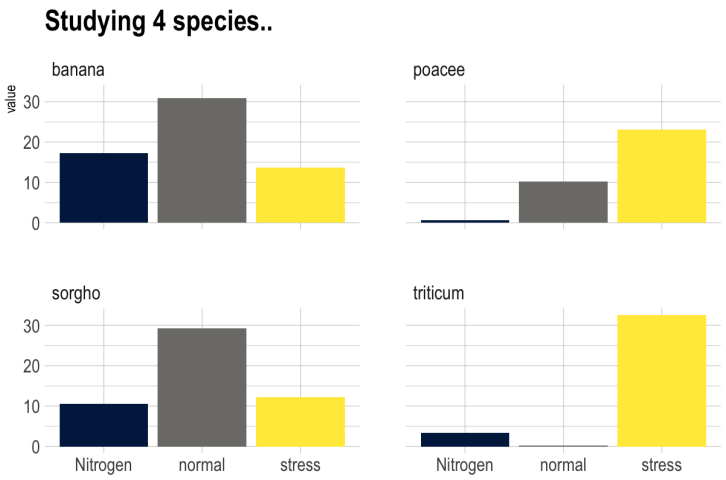
Price

Purchase decision Price	Absolute	Relative
1	23	11.44%
2	51	25.37%
3	47	23.38%
4	41	20.40%
5	39	19.40%





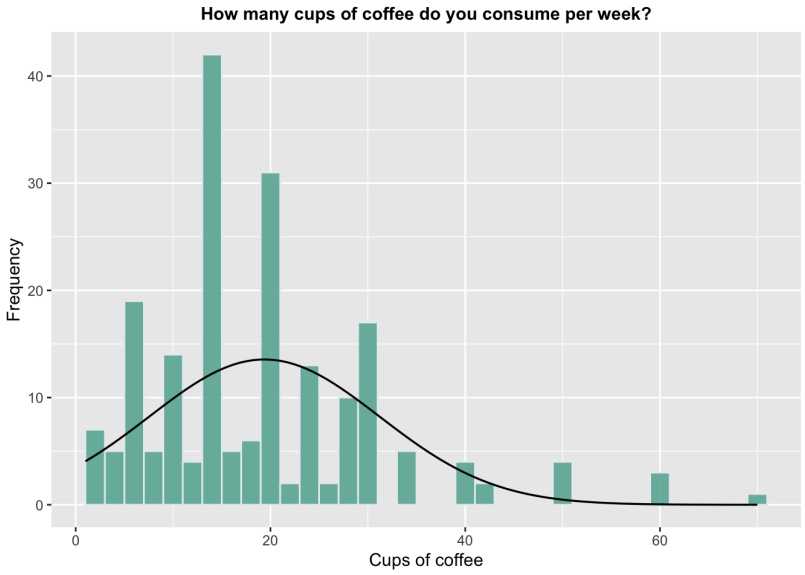

```
Loading required package: viridisLite
```


Studying 4 species..
banana
poacee
sorgho
triticum
value
30
20
10
0
Nitrogen
normal
stress

Univariate descriptions - Numerical variables

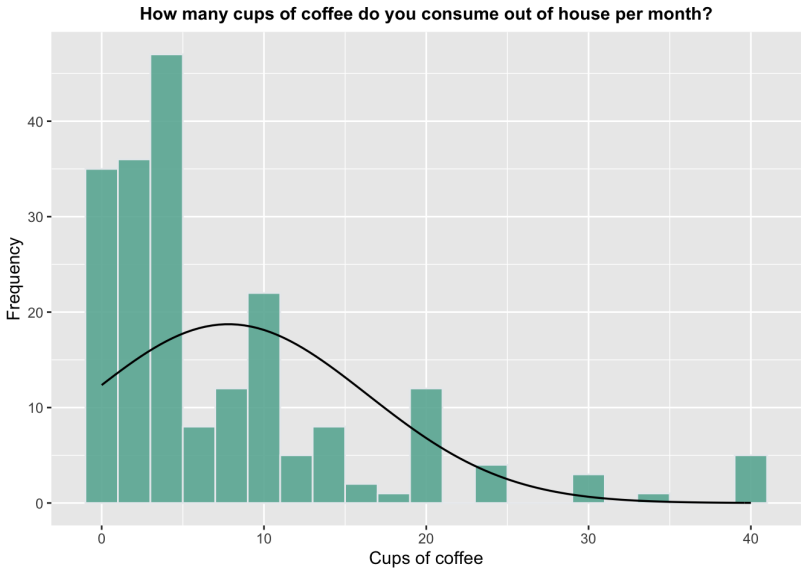
Amount coffe consumed weekly

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   12.00   17.00   19.38   25.00   70.00
```



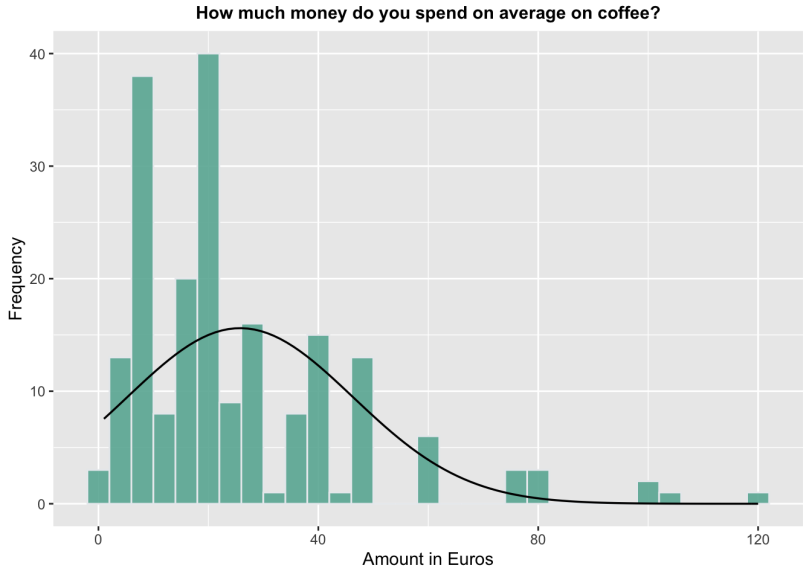
Amount per month out of house

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   5.000   7.811  10.000  40.000
```



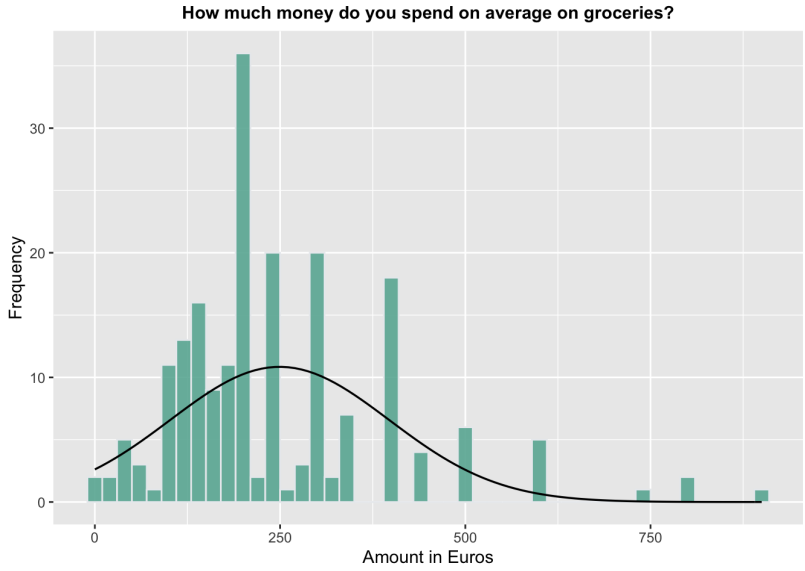
Money coffee

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   10.00   20.00   25.77   35.00  120.00
```



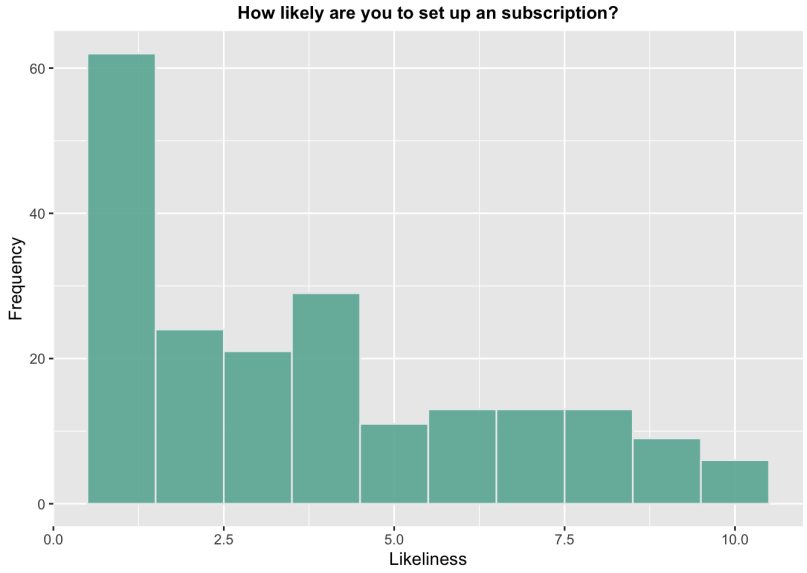
Money groceries

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   150.0   200.0   249.5   300.0   900.0 
```



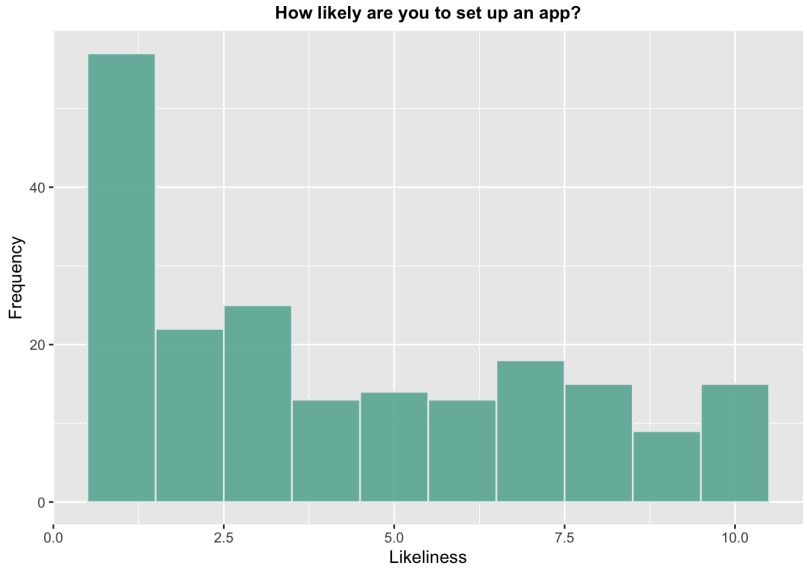
Subscription likely

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   3.000   3.771   6.000  10.000 
```

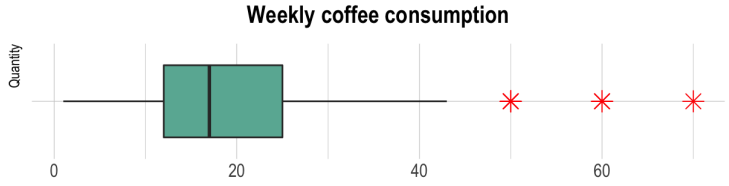


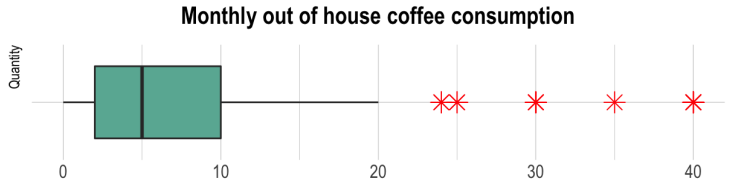
App likely

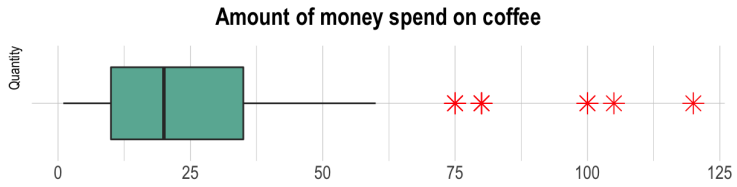
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   1.000   3.000   4.244   7.000  10.000
```

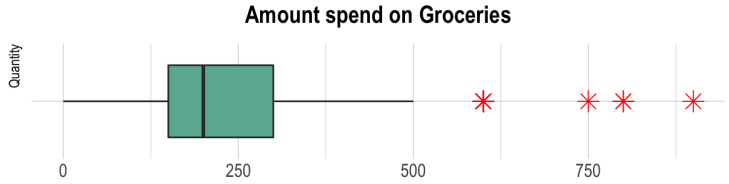


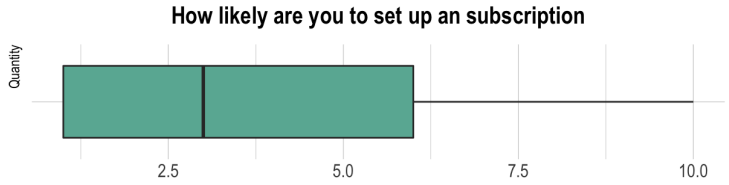
Boxplots

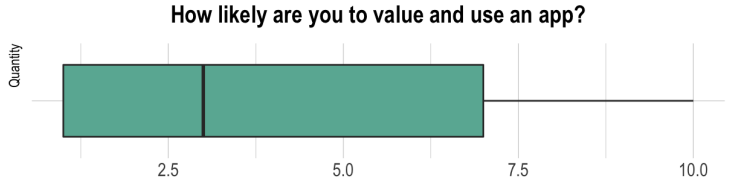

Weekly coffee consumption
Quantity
0
20
40
60


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









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 = 230.83, df = 132, p-value = 0.0000002273


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

```
data:  AmountWeek and AgeCategory
X-squared = 230.83, df = NA, p-value = 0.03194
```

Education - Amount coffee drank

```
    Pearson's Chi-squared test

data:  AmountWeek and Education
X-squared = 224.72, df = 165, p-value = 0.001378


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

data:  AmountWeek and Education
X-squared = 224.72, df = NA, p-value = 0.07784
```

Gender - Amount coffee drank

```
    Pearson's Chi-squared test

data:  AmountWeek and Gender
X-squared = 69.007, df = 66, p-value = 0.3761


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

data:  AmountWeek and Gender
X-squared = 69.007, df = NA, p-value = 0.3433
```

Home - Amount coffee drank

```
    Pearson's Chi-squared test

data:  AmountWeek and Home
X-squared = 60.127, df = 66, p-value = 0.6804


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

data:  AmountWeek and Home
X-squared = 60.127, df = NA, p-value = 0.6567
```

App - Age

```
    Pearson's Chi-squared test

data:  App_Likely and AgeCategory
X-squared = 52.761, df = 36, p-value = 0.0353


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

data:  App_Likely and AgeCategory
X-squared = 52.761, df = NA, p-value = 0.02994
```

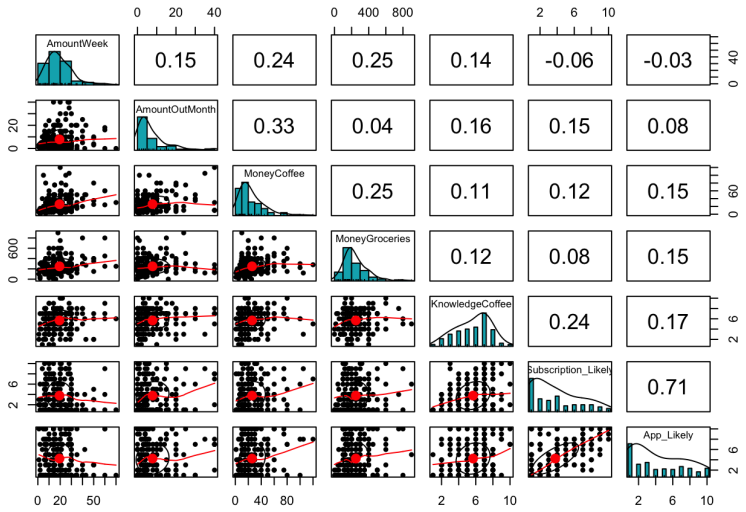
Coffee knowledge - Age

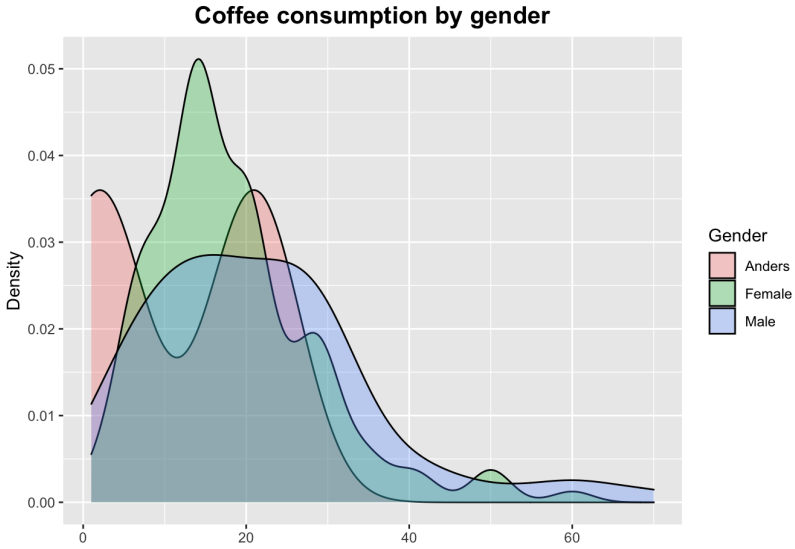
```
    Pearson's Chi-squared test

data:  KnowledgeCoffee and AgeCategory
X-squared = 104.25, df = 36, p-value = 0.00000001471


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

data:  KnowledgeCoffee and AgeCategory
X-squared = 104.25, df = NA, p-value = 0.003992
```

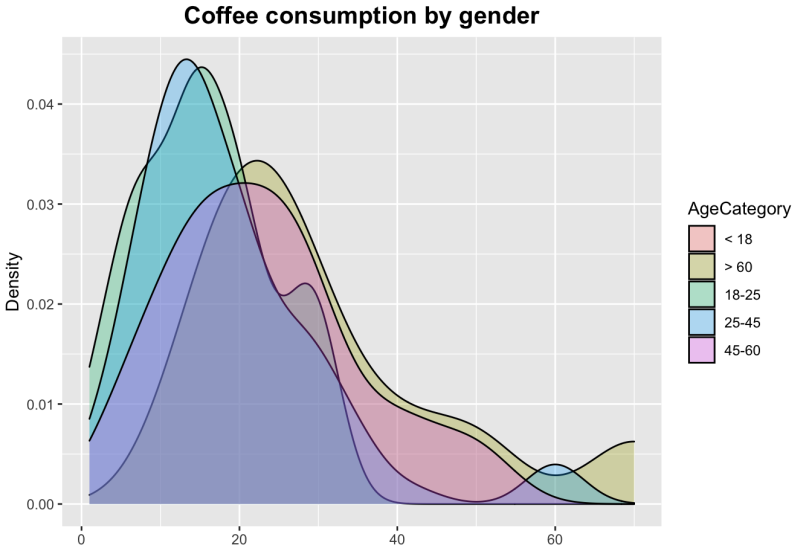
Coffee knowledge - Purchase location

```
    Pearson's Chi-squared test

data:  KnowledgeCoffee and PurchaseLocation
X-squared = 50.617, df = 72, p-value = 0.9738


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

data:  KnowledgeCoffee and PurchaseLocation
X-squared = 50.617, df = NA, p-value = 0.9281
```

Relationships






Coffee consumption by gender
Density
0.04
0.03
0.02
0.01
0.00
0
20
40
60
AgeCategory
< 18
> 60
18-25
25-45
45-60

Regressions

Incl categorical variables as dummies

Cooks distance –> outliers

Data problems