Data Acquisition and Survey Methods Report

Ivana Dasović Claudia Kößldorfer Fani Sentinella-Jerbić

2023-06-23

Contents

Introduction

In today's fast-paced world, the decision of whether to cook or order food has become more and more common, especially with students. Balancing academic demands, social life, and personal well-being can leave students stuck in trade-offs between convenience, nutrition, and entertainment. To gain a comprehensive understanding of their habits, we conducted a survey investigating their frequency of cooking and ordering food, as well as the underlying motivations and time commitments associated.

Survey Design

In our survey we attempted to capture the intricacies of student behavior and preferences around cooking and ordering food. To clarify the terminology used, we define cooking as the act of preparing a meal oneself, excluding the reheating of pre-prepared dishes. Ordering food, on the other hand, refers to the process of obtaining meals cooked by someone else, such as through food delivery services or dining out. We present all the questions and the motivation behind choosing each below.

Question 1: *How often do you cook or order food in two weeks?* [2 number inputs]

This question was supposed to give us insight into proportion of cooking versus ordering food events. We chose a time frame of two weeks because it is not too far in the past that the participants can't remember but also not too short to account for specifics of certain weeks in participants' lives (such as exams, trips, or formal events).

Question 2: *When you order food what is your most common reason?* [multiple choice]

- *Lack of time for cooking*
- *Want specific food/cuisine*
- *Lack of motivation for cooking ("laziness")*
- *As a reward*
- *Eating as a group*
- *Lack of cooking skills*
- *Other*

We sought to understand the motivations behind ordering food. Is it a lack of time or more of a preference? The survey presented a list of potential reasons, and an option for other reasons not listed.

Question 3: *When you cook how much time do you spend on average? (in minutes)* [number input] Lastly, we ask for the average time commitment associated with cooking. This question provides valuable information on the time allocation that students are willing to invest in cooking, which can help contextualize their choices between cooking and ordering food.

Other than these domain-specific questions, our survey included general demographic questions about sex, age, and study program.

Results Preprocessing

This section focuses on the preprocessing steps undertaken to clean and prepare the collected survey responses for subsequent analysis.

Special remark to note here is that our first survey question was implemented with only one number input instead of two, which probably made this question confusing for our respondents and thus further analysis of the responses for this question may not be as interpretable as we had anticipated.

Cleaning and Formatting

Upon reviewing the data, we observed variations in the input format for academic programs. To ensure consistency, our first step involved unifying these entries into a standardized format.

Next, we checked if all the survey responses were reasonable or if there were any anomalies or outliers. During this review, we came across an outlier in the first question, where a respondent had entered the value of 150. Interpreting this response as cooking or ordering food more than ten times a day within a two-week period seemed improbable, given the range of other answers. It appeared to be a potential typographical error, and a value of 15 seemed more consistent with the other responses. Thus, we made the adjustment accordingly. For the second question, it was single-choice, so there couldn't be any deviation from the format. In the case of the third question, one respondent provided a response in the form of a range - '30-40'. To handle this situation, we calculated the mean value of the range and used it as the representative answer.

Lastly we changed some of the data types to more sensible ones. Gender, academic program and answers to the second question should be factors, while responses to the third question should be an integer.

Exploration

To obtain a summary of the data distribution, we used the function *tbl_summary*. This allowed us to calculate descriptive statistics for various types of variables within the dataset.

```
data %>%
  tbl_summary(
    type = all_continuous() ~ "continuous2",
    statistic = list(
      all_continuous() ~ c("{median} ({p25}, {p75})", "{min}, {max}"),
      all_categorical() ~ "{n} / {N} ({p}%)")) %>%
  bold_labels()
```

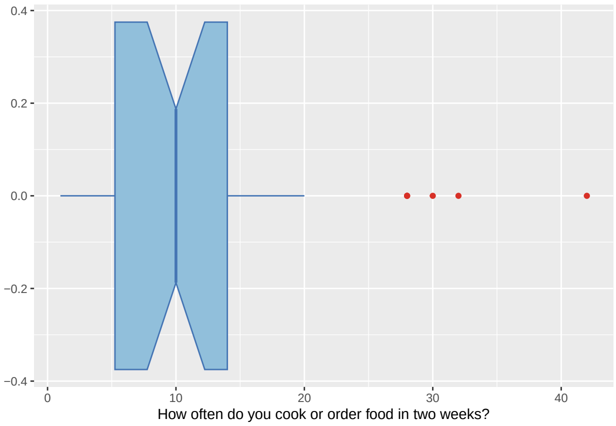
Characteristic	**N = 38**
Gender	
female	16 / 38 (42%)
male	22 / 38 (58%)
Age	
Median (IQR)	24 (23, 27)
Range	21, 46
Academic.Program	
Business Informatics / BSc / TU Wien	2 / 38 (5.3%)
Data Science / MSc / TU Wien	32 / 38 (84%)
Data Science / MSc / University of Zagreb	1 / 38 (2.6%)
Erasmus student, Statistic	1 / 38 (2.6%)
MSc	1 / 38 (2.6%)
Statistik und Wirtschaftsmathematik / BSc / TU Wien	1 / 38 (2.6%)
Antwort.1	
Median (IQR)	10 (5, 14)
Range	1, 42
Antwort.2	
As a reward	1 / 38 (2.6%)
Eating as a group	7 / 38 (18%)
Lack of motivation for cooking ("laziness")	13 / 38 (34%)
Lack of time for cooking	8 / 38 (21%)
Other	1 / 38 (2.6%)
Want specific food/cuisine	8 / 38 (21%)
Antwort.3	
Median (IQR)	43 (30, 48)
Range	20, 90

The table provides a snapshot of the data. We can see there are 16 female and 22 male students included in the analysis, with age ranging from 21 to 46, with a median age of 24. Majority of the students are pursuing studies in Data Science at the TU Wien.

The table also presents the statistics related to our survey questions. However, visual plots are available below for a more comprehensive understanding. The plots additionally reveal outliers, specifically in first and third question responses.

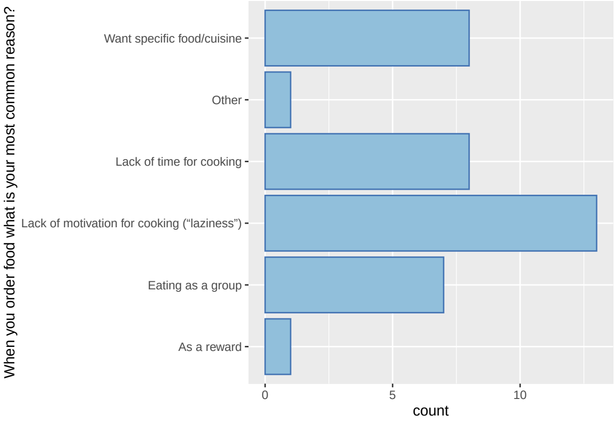
Question 1: *How often do you cook or order food in two weeks?*

```
ggplot(data) +
  geom_boxplot(aes(x=Antwort.1),notch=TRUE,outlier.colour="#d73027",fill='#91bfdb', color="#4575b4") +
  xlab("How often do you cook or order food in two weeks?")
```



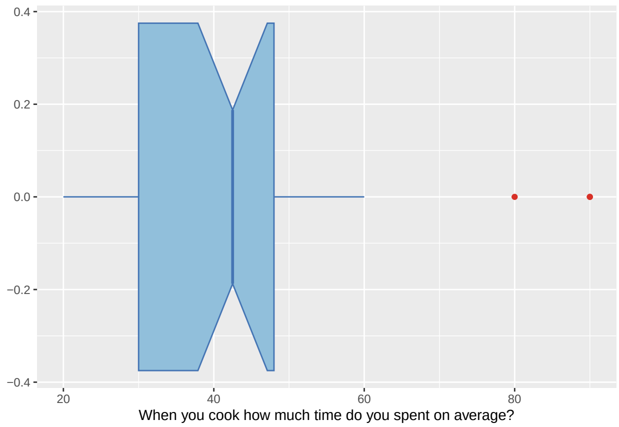
Question 2: *When you order food what is your most common reason?*

```
ggplot(data) +
  geom_bar(aes(y = Antwort.2), fill='#91bfdb', color="#4575b4") +
  ylab("When you order food what is your most common reason?")
```



Question 3: *When you cook how much time do you spend on average? (in minutes)*

```
ggplot(data) +
  geom_boxplot(aes(x=Antwort.3),notch=TRUE,outlier.colour="#d73027",fill='#91bfdb', color="#4575b4") +
  xlab("When you cook how much time do you spent on average?")
```


0.4
0.2
0.0
−0.2
−0.4
20
40
60
80
When you cook how much time do you spent on average?

Research questions

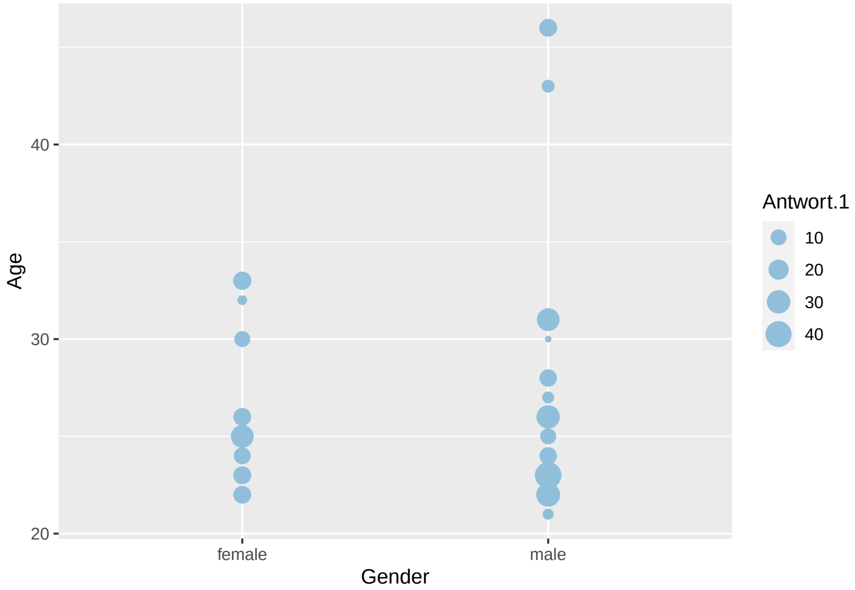
Research question 1: Age and gender affect how often people cook/order food.

We wanted to investigate the relationship between and the possible impact of age, gender, and the frequency of cooking or ordering food as these factors could influence peoples eating habits.

Explorative Data Analysis

While the data has an almost even split when it comes to gender, there is a noticeable clustering of ages among the students. This can be seen when exploring the relevant variables. In the following plot the responses to the first question are shown, with the size of the data points along the age axis corresponding to the answers given by females or males. No discernible pattern is visible, but the majority of students are below the age of 25, making their data points challenging to interpret.

```
ggplot(data, aes(x=Gender, y=Age)) +
  geom_point(aes(size=Antwort.1), color="#91bfdb")
```



Based on this, it is clear that the data needs to be further analyzed. By splitting the data into age groups of approximately equal sizes, the effect of age can be better visualized and also paired with the gender variable.

Descriptive Inference

tbl_summary was used to create two tables; one for female and one for male students. These tables show for each age group the size of the groups in each subset and important summary statistics for the answer to our first survey question.

```r
age_sequence = c(21,22,23,28,46)

data_1 <- data %>%
  select(Age, Gender,Antwort.1) %>%
  mutate(Age_Groups = cut(Age,age_sequence, include.lowest=TRUE))

data_1 %>%
  filter(Gender == "female")%>%
  select(Age_Groups, Antwort.1) %>%
  tbl_summary(by = Age_Groups,
              type = all_continuous() ~ "continuous2",
              statistic = all_continuous() ~ c("{median} ({p25}, {p75})", "{min}, {max}")) %>%
  as_gt() %>%
  tab_header(title = md("Antwort.1 by Age Groups when Gender is female"))

data_1 %>%
  filter(Gender == "male")%>%
  select(Age_Groups, Antwort.1) %>%
  tbl_summary(by = Age_Groups,
              type = all_continuous() ~ "continuous2",
              statistic = all_continuous() ~ c("{median} ({p25}, {p75})", "{min}, {max}")) %>%
  as_gt() %>%
  tab_header(title = md("Antwort.1 by Age Groups when Gender is male"))
```

Antwort.1 by Age Groups when Gender is female

Characteristic	**[21,22]**, N = 4	**(22,23]**, N = 4	**(23,28]**, N = 4	**(28,46]**, N = 4
Antwort.1				
Median (IQR)	11.5 (8.8, 13.3)	10.5 (5.8, 14.0)	13.0 (12.0, 17.5)	9.0 (6.5, 11.3)
Range	5.0, 14.0	2.0, 14.0	12.0, 28.0	2.0, 15.0

Antwort.1 by Age Groups when Gender is male

Characteristic	**[21,22]**, N = 4	**(22,23]**, N = 6	**(23,28]**, N = 8	**(28,46]**, N = 4
Antwort.1				
Median (IQR)	24 (16, 29)	8 (5, 10)	10 (7, 13)	10 (4, 18)
Range	3, 32	4, 42	4, 30	1, 28

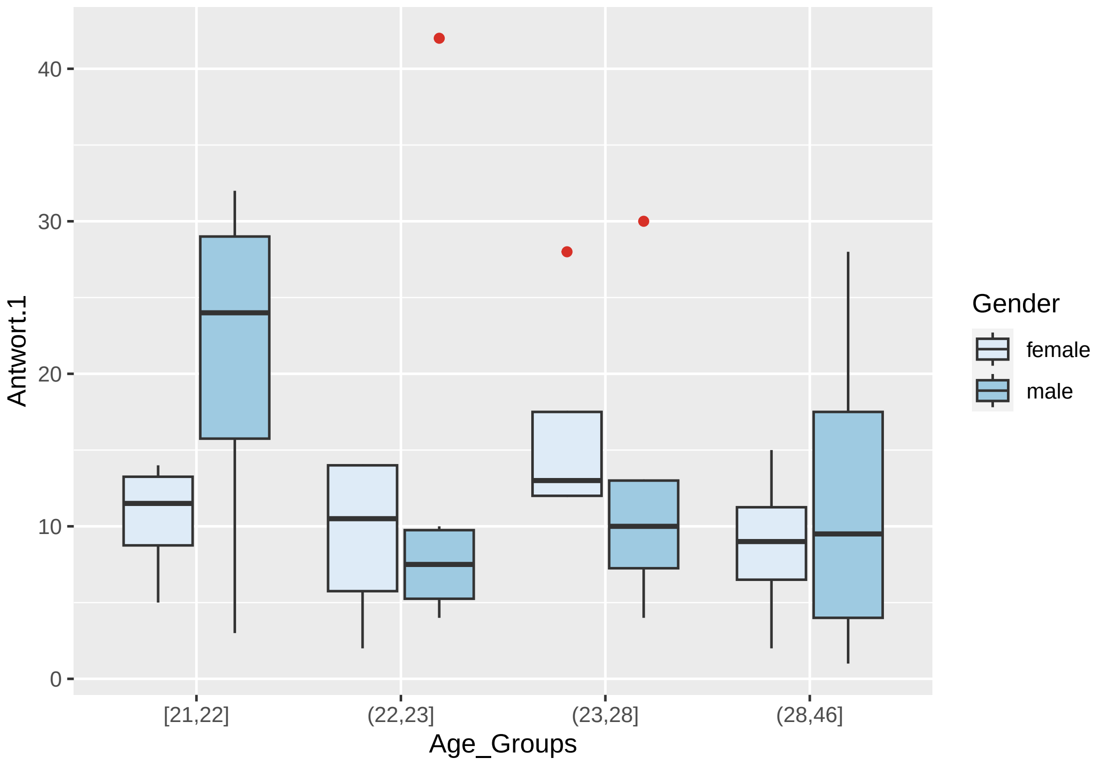
From the table for female students no clear trend can be identified. All the age groups are of the same size and their medians are close in range of each other. Though the third age group (23,28] covers the largest range of answer values for question one.

In contrast there is a very clear outlier in the data of the male students, with the first age group having a median that is more than double those of the other age groups. While other age groups cover similar or even larger value ranges, their medians are in the lowest third of their total range.

All these observations are supported in the following boxplot. The similarity between the female age groups is clear as is the extreme disparity between the first male age group and the rest. Furthermore, in the plots outliers which led to very large ranges of value but far smaller medians are also visualized.

From all this we could assume, due to only one age group of one gender showing a clear trend that our assumption of gender and age affecting how often people cook/order food might not be supported.

```
ggplot(data_1) +
  geom_boxplot(aes(x=Age_Groups,y=Antwort.1,fill=Gender),outlier.colour="#d73027") +
  scale_fill_brewer(palette = "Blues")
```



Analytic Inference

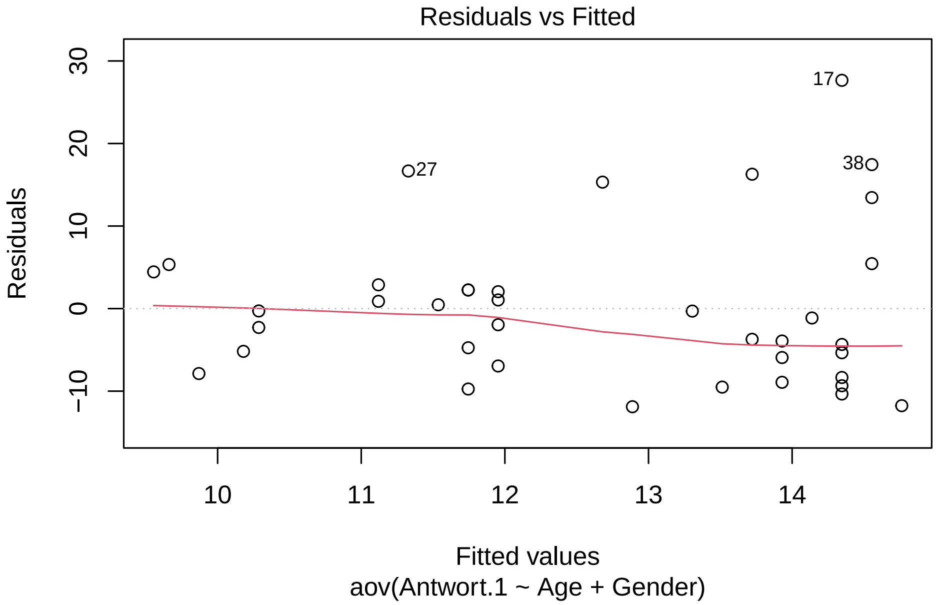
We tested with a two-sided ANOVA test, to analyze the difference between the means of both gender and age. When conducting such a test, the p-values quantify the likelihood of observing the data or more extreme results under the assumption that the null hypothesis is true, meaning there is no significant difference between the groups compared.

```
aovTest <- aov(Antwort.1 ~ Age + Gender, data = data)
summary(aovTest)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Age          1     38   37.54   0.408  0.527
## Gender       1     62   62.14   0.675  0.417
## Residuals   35   3221   92.03
```

From the summary of the test we can see that for age the p-value is 0.527 and for gender it is 0.417. These values clearly far surpass the common alpha level of 0.05 which leads to us not being able to reject the null hypothesis of this test.

```
plot(aovTest, 1)
```



Even though we have some outliers in the residuals, overall we have to assume that no statistical significant difference in the means across age and gender can be proven with our data. This would suggest that the differences we could see in the data for the youngest male age group could be simply random chance or actually the effect of factors we didn't account for.

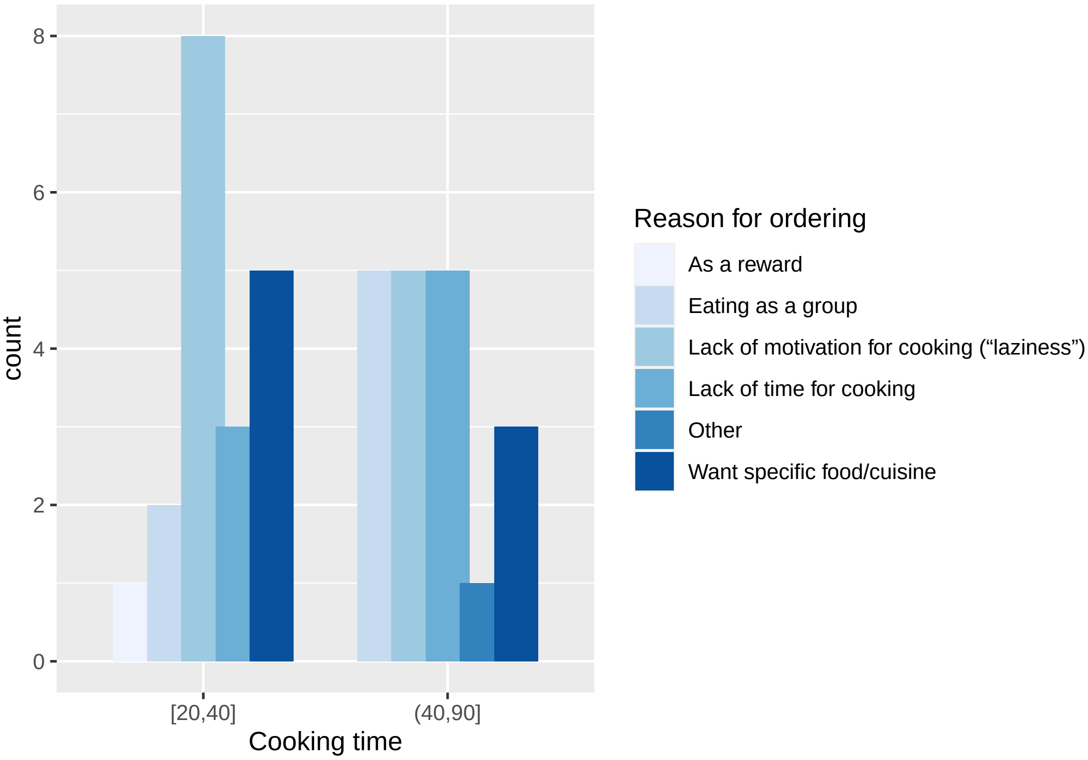
Research question 2: Reason to order food depends on how much time someone on average spends cooking.

Through this research question we aim to understand the relationship between the time spent cooking and the reasons behind ordering food.

Exploratory Data Analysis

Presented in the form of a bar plot, we can observe and compare the factors that influence the decision to order food between two distinct groups based on their average cooking time. We approached this with splitting respondents to ones who cook for shorter time periods and those who cook for longer time periods. The first group comprises of individuals who spend between 20 to 40 minutes on cooking, while the second group consists of those who dedicate 40 to 90 minutes to cooking.

```r
antwort3_sequence = c(20,40,90)
A3_Groups = cut(data$Antwort.3,antwort3_sequence, include.lowest=TRUE)

ggplot(data, aes(A3_Groups, after_stat(count))) +
  geom_bar(aes(fill = Antwort.2), position = position_dodge(width=0.7)) +
  scale_fill_brewer(palette = "Blues") +
  labs(fill = "Reason for ordering", x="Cooking time")
```



By analyzing the bar plot, we can gain insights into the reasons that motivate each group to order food instead of preparing meals at home. One notable observation is that individuals who spend less than 40 minutes on cooking typically attribute their choice to order food to a lack of motivation, rather than a shortage of time. In other words, time constraints do not appear to be the primary factor for this group.

Another common reason for this group is the craving of a specific cuisine which the respondent may not be able to cook themselves.

On the other hand, when examining the second group, no single reason emerges as a dominant factor influencing their decision. The bar plot does not exhibit a pronounced precedence of one reason over others within this group. It suggests that individuals who spend between 40 to 90 minutes on cooking have a more diverse range of motivations for opting to order food.

Descriptive Inference

Continuing with the same idea, we investigate summary statistics through tables and box plots of average cooking time by reason for ordering.

The observation reveals that the most commonly selected response, which is a lack of motivation, corresponds to the highest average cooking time.

Following this, in decreasing order, we have "other" reasons, although it is worth noting that only one person chose this option. Same is the case for rewarding oneself, however it has the lowest average cooking time.

The category of eating as a group exhibits the lowest standard deviation among the well-represented categories.

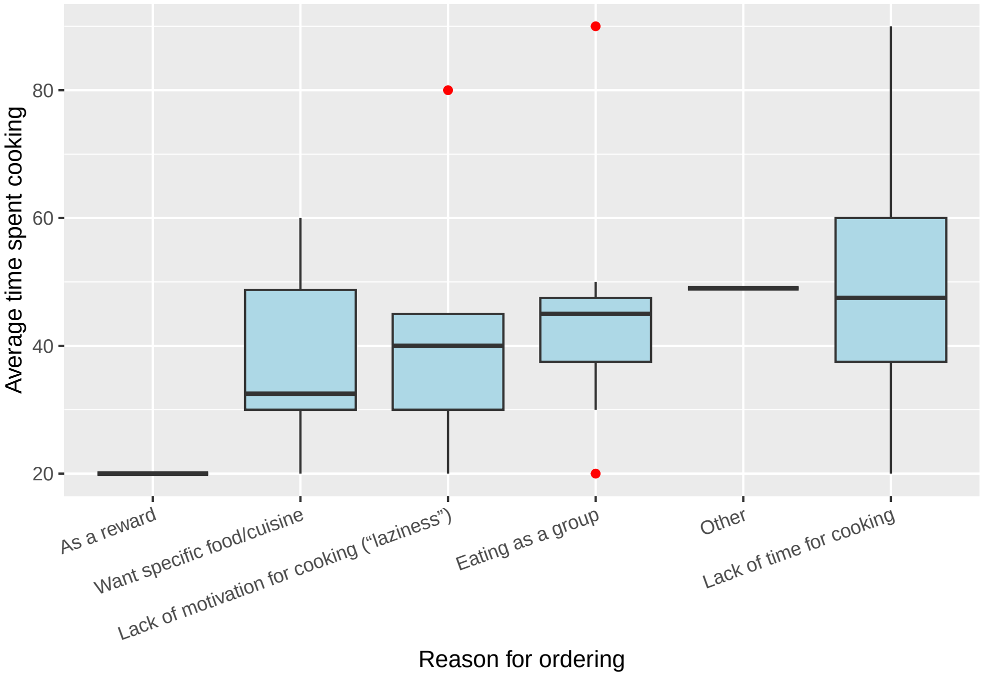
Finally, we have the categories of lack of motivation and cravings for specific cuisine.

```r
data_2 <- data %>%
  mutate(Antwort.3_Groups = A3_Groups)

data_2_stat <- data_2 %>% group_by(Antwort.2) %>%
  summarise(
    Count = n(),
    Mean = round(mean(Antwort.3, na.rm = TRUE), digits = 2),
    Median = median(Antwort.3),
    SD = round(sd(Antwort.3, na.rm = TRUE), digits = 2),
    Min = min(Antwort.3),
    Max = max(Antwort.3)
  ) %>%
  arrange(Mean)

gt(data_2_stat) %>%
  tab_header(title = "Antwort.3 Statistics dependent on Antwort.2") %>%
  tab_spanner(label = "Antwort.3 Statistics",
              columns = (!contains("Antwort") & !contains("Count")))
```

Antwort.3 Statistics dependent on Antwort.2

		Antwort.3 Statistics				
Antwort.2	Count	Mean	Median	SD	Min	Max
As a reward	1	20.00	20.0	NA	20	20
Want specific food/cuisine	8	38.75	32.5	14.82	20	60
Lack of motivation for cooking ("laziness")	13	39.62	40.0	14.50	20	80
Eating as a group	7	46.43	45.0	21.93	20	90
Other	1	49.00	49.0	NA	49	49
Lack of time for cooking	8	49.38	47.5	21.45	20	90

```
ggplot(data, aes(x = reorder(Antwort.2, Antwort.3), y = Antwort.3)) +
  geom_boxplot(outlier.colour = "red", fill="lightblue") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
  labs(x = "Reason for ordering", y = "Average time spent cooking")
```



Analytic Inference

To assess this research question we employed the Pearson's Chi-squared test which used to determine whether there is a significant association between two categorical variables.

```
chisq.test(table(data_2$Antwort.2,data_2$Antwort.3_Groups),
           correct = FALSE)
```

```
## Warning in chisq.test(table(data_2$Antwort.2, data_2$Antwort.3_Groups), :
## Chi-Quadrat-Approximation kann inkorrekt sein
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data_2$Antwort.2, data_2$Antwort.3_Groups)
## X-squared = 4.978, df = 5, p-value = 0.4186
```

The results of the Pearson's Chi-squared test indicate that there is no significant association between the reason for ordering and average time spent cooking. The test statistic is calculated as 4.978 with a degree of freedom 5. Based on the p-value of 0.4186, which exceeds the significance level of 0.05, we fail to reject the null hypothesis. This implies that there is insufficient evidence to suggest a meaningful relationship. The reason for this may be the underrepresentation of certain reason groups.

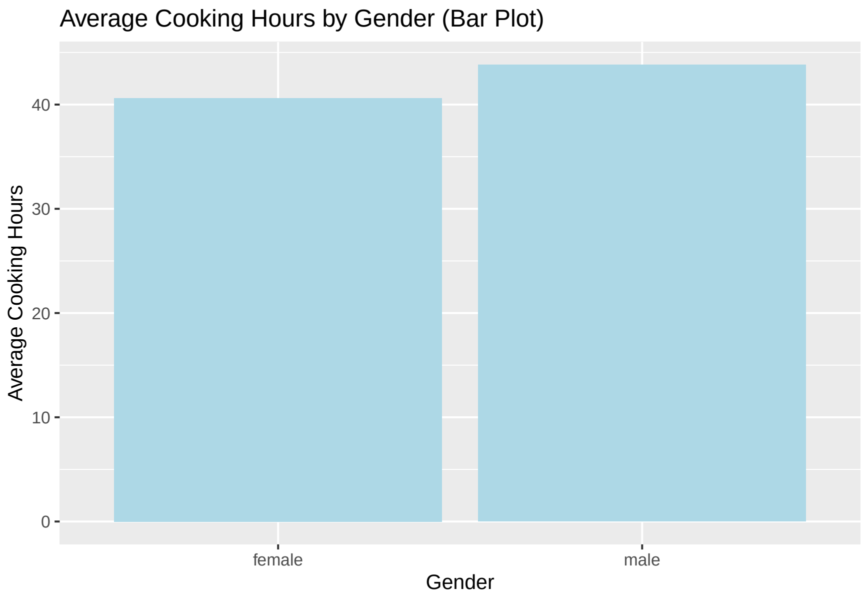
Research question 3: On average females spend more time on cooking than males.

With this research question our objective was to examine the relationship between gender and the potential influence on cooking time.

Exploratory Data Analysis

As evidenced by the bar plot provided, the average cooking hours differ between the genders. However, it is important to note that these differences in means do not necessarily indicate a significant distinction. In order to determine if the disparities observed are statistically significant, further analysis and hypothesis testing would be required.

```
data_3_summary <- data %>%
  group_by(Gender) %>%
  summarize(
    Mean = round(mean(Antwort.3, na.rm = TRUE), digits = 2))

ggplot(data_3_summary, aes(x = Gender, y = Mean)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  labs(x = "Gender",
       y = "Average Cooking Hours",
       title = "Average Cooking Hours by Gender (Bar Plot)")
```



Descriptive Inference

As evident from the tabulated information, the average cooking time data categorized by gender reveals interesting insights. This data, which is also visually depicted in the accompanying boxplots, sheds light on the distribution, central tendencies, and variabilities associated with cooking times for females and males.

The median cooking time for females is recorded at 38 minutes, accompanied by an interquartile range spanning from 30 to 45 minutes. In contrast, males exhibit a slightly higher median cooking time of 45 minutes, with an interquartile range ranging from 30 to 50 minutes. Notably, both genders share an identical range of cooking times, encompassing values from 20 to 90 minutes.

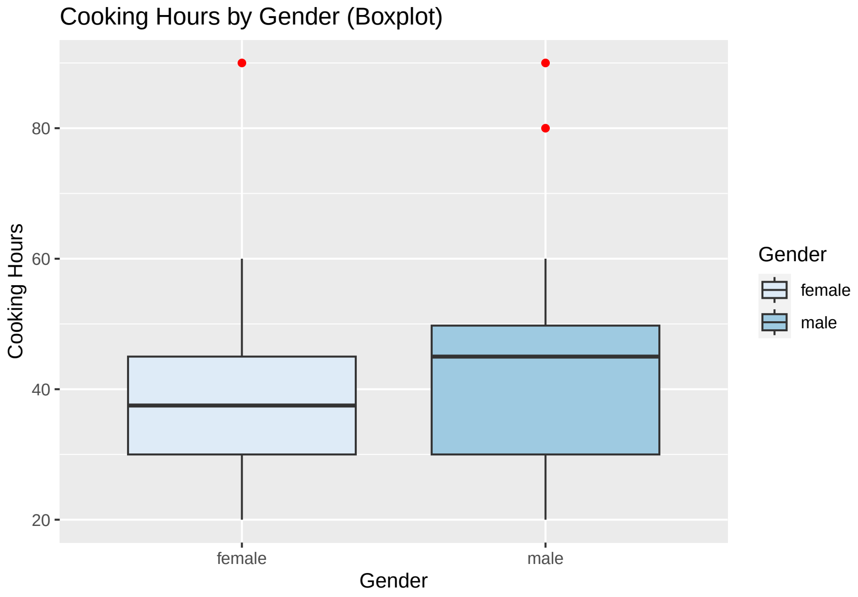
By looking at the boxplots, we can understand how the cooking time is distributed for each gender and identify any noticeable differences or unusual values. This helps us get a better idea of how cooking time varies between females and males.

```r
data %>%
  select(Antwort.3, Gender) %>%
  tbl_summary(by = Gender,
    type = all_continuous() ~ "continuous2",
    statistic = all_continuous() ~ c("{median} ({p25}, {p75})", "{min}, {max}"))%>%
  as_gt() %>%
  tab_header(
    title = md("Antwort.3 by Gender"),
  )
```

Antwort.3 by Gender

Characteristic	**female**, N = 16	**male**, N = 22
Antwort.3		
Median (IQR)	38 (30, 45)	45 (30, 50)
Range	20, 90	20, 90

```r
ggplot(data, aes(x = Gender, y = Antwort.3, fill= Gender)) +
  geom_boxplot(outlier.colour = "red") +
  scale_fill_brewer(palette = "Blues") +
  labs(x = "Gender", y = "Cooking Hours", title = "Cooking Hours by Gender (Boxplot)")
```



Analytic Inference

To determine if there is a significant difference in the mean time spent on cooking between males and females, we conducted a t-test. This statistical test allows us to assess whether the observed means for both genders are significantly different from each other.

```
data_3 <- data %>% rename(cooking_hours = Antwort.3)

data_females <- subset(data_3, Gender == "female")
data_males <- subset(data_3, Gender == "male")

t.test(data_females$cooking_hours, data_males$cooking_hours, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  data_females$cooking_hours and data_males$cooking_hours
## t = -0.55403, df = 33.721, p-value = 0.7084
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -12.94118       Inf
## sample estimates:
## mean of x mean of y
##  40.62500  43.81818
```

The results of the Welch two-sample t-test indicate that there is no significant difference in the mean time spent on cooking between females and males. The calculated t-value is -0.55403, with a corresponding p-value of 0.7084. The calculated p-value is greater than the significance level of 0.05. This suggests that the observed difference in means is not statistically significant, and any observed differences in cooking hours between the two genders may be attributed to chance or random variability in the sample, rather than a true difference at the population level. The sample estimates show that the mean time spent on cooking for females is 40.62500, while for males, it is 43.81818. The failure to reject the null hypothesis indicates that the alternative hypothesis, which states that there is a true difference in means greater than 0, is not supported.