

# Political Science Research Methods Lab Assignments

Kelsey Quigg

Fall 2023

## Part One: Data Cleaning

Load packages and data:

```
#install.packages("pacman")
pacman::p_load(dplyr, ggplot2, openxlsx, tidyverse)
setwd("/Users/kelseyquigg/Documents/PortfolioWebsite/Projects")
df <- read.xlsx("Week4DataCopy.xlsx")
head(df, 20)
```

##	Sex	Age	Hometown	FavoriteMeat	FavoriteSauce
## 1	Female	21	South Carolina	Pork Ribs	Dry Rub
## 2	Female	18	Piedmont	Pork Ribs	Other
## 3	Male	20	Elsewhere	Pork Ribs	Dry Rub
## 4	Female	21	South Carolina	Beef Ribs	Other
## 5	Male	22	Eastern or Central NC	Beef Ribs	Eastern style (with no tomato)
## 6	Male	44	Elsewhere	Poultry	Korean Style
## 7	Female	22	Piedmont	Pork Ribs	Eastern style (with no tomato)
## 8	Female	20	Piedmont	Poultry	Western style (with tomato)
## 9	Female	21	Eastern or Central NC	Pulled Pork	Eastern style (with no tomato)
## 10	Female	21	South Carolina	Pulled Pork	Korean Style
## 11	Male	21	Piedmont	Pulled Pork	Eastern style (with no tomato)
## 12	Male	25	Eastern or Central NC	Poultry	Dry Rub
## 13	Female	43	Elsewhere	Poultry	Kansas style (with molasses)
## 14	Female	22	Piedmont	Beef ribs	Eastern style (with no tomato)
## 15	Female	23	Eastern or Central NC	Pulled Pork	Eastern style (with no tomato)
## 16	Male	46	Eastern or Central NC	Pulled Pork	Eastern style (with no tomato)
## 17	Male	23	Eastern or Central NC	Beef ribs	Western style (with tomato)
## 18	Male	21	Piedmont	Pork Ribs	Kansas style (with molasses)
## 19	Male	22	Eastern or Central NC	Pulled Pork	South Carolina Mustard
## 20	Female	17	Piedmont	Poultry	Dry Rub
##	Sweetness	FavoriteSide	MinutesDriving	SandwichPrice	DinnerPlatePrice
## 1	4	Collard greens	10	20	22
## 2	4	Hush puppies	20	12	25
## 3	3	Collard greens	60	10	20
## 4	5	Baked beans	35	13	16
## 5	3	Other	20	12	16
## 6	3	Other	60	10	15
## 7	3	Baked beans	40	15	25
## 8	3	Fries	20	16	22
## 9	2	Collard greens	15	14	16

```
## 10      4      Other      20      12      15
## 11      3      Other      45      10      15
## 12      4      Other      30      15      18
## 13      3      Baked Beans  45      12      18
## 14      2      Baked Beans  60      15      20
## 15      2      Baked Beans  60      15      23
## 16      1 Collard greens  120     20      40
## 17      5      Fries      60      15      25
## 18      3      Other      30      16      18
## 19      2      Coleslaw    90      25      35
## 20      1      Fries      30      15      20
##      RibsPrice
## 1      24
## 2      25
## 3      25
## 4      15
## 5      18
## 6      18
## 7      35
## 8      22
## 9      22
## 10     19
## 11     25
## 12     25
## 13     24
## 14     20
## 15     27
## 16     30
## 17     25
## 18     26
## 19     35
## 20     30
```

Add observation numbers:

```
df <- df %>%
  rowid_to_column(var = "Observation")
```

## Recoding Categorical Variables to Numbers

**Gender** Convert each category to its respective number based on the survey codebook:

```
unique(df$Sex)
```

```
## [1] "Female"    "Male"      "Other"     "They/them" "Demifluid" NA
## [7] "female"    "male"
```

```
df <- df %>%
  mutate(Sex =
    case_match(
      Sex,
      c("Male", "male") ~ "1",
```

```

      c("Female", "female") ~ "2",
      c("Other", "They/them", "Demifluid") ~ "3",
      .default = Sex))
df$Sex <- as.numeric(as.character(df$Sex))

```

```
unique(df$Hometown)
```

## Hometown

```

## [1] "South Carolina"      "Piedmont"
## [3] "Elsewhere"           "Eastern or Central NC"
## [5] "Texas"               "Western NC"
## [7] "Eastern or central NC" "Eastern or Central Nc"
## [9] "Eastern/Central NC"  NA
## [11] "Tennessee"          "Piedmount"
## [13] "Eastern or Central Carolina" "Tennessee"
## [15] "Peidmont"            "Somewhere else"
## [17] "Tennesse"

```

```

df <- df %>%
  mutate(Hometown =
    case_match(
      Hometown,
      c("Eastern or Central NC", "Eastern or central NC", "Eastern or Central Nc",
        "Eastern/Central NC", "Eastern or Central Carolina") ~ "1",
      c("Piedmont", "Piedmount", "Peidmont") ~ "2",
      "Western NC" ~ "3",
      "South Carolina" ~ "4",
      "Texas" ~ "5",
      c("Tennessee", "Tennessee", "Tennesse") ~ "6",
      c("Elsewhere", "Somewhere else") ~ "7",
      .default = Hometown))
df$Hometown <- as.numeric(as.character(df$Hometown))

```

```
unique(df$FavoriteMeat)
```

## Favorite Meat

```

## [1] "Pork Ribs"      "Beef Ribs"      "Poultry"        "Pulled Pork"
## [5] "Beef ribs"      "Pulled pork"     "Beef Brisket"   "pork ribs"
## [9] "beef ribs"      "pulled pork"     "poultry"         "beef brisket"
## [13] "Pulled Ribs"    "Beef brisket"    "Pork ribs"       "Beef Briskets"
## [17] NA

```

```
df <- df %>%
  mutate(FavoriteMeat =
    case_match(
      FavoriteMeat,
      c("Pulled Pork", "Pulled pork", "pulled pork") ~ "1",
      c("Pork Ribs", "pork ribs", "Pork ribs", "Pulled Ribs") ~ "2",
      c("Beef Brisket", "beef brisket", "Beef brisket", "Beef Briskets") ~ "3",
      c("Beef Ribs", "Beef ribs", "beef ribs") ~ "4",
      c("Poultry", "poultry") ~ "5",
      .default = FavoriteMeat))
df$FavoriteMeat <- as.numeric(as.character(df$FavoriteMeat))
```

```
unique(df$FavoriteSauce)
```

## Favorite Sauce

```
## [1] "Dry Rub"
## [2] "Other"
## [3] "Eastern style (with no tomato)"
## [4] "Korean Style"
## [5] "Western style (with tomato)"
## [6] "Kansas style (with molasses)"
## [7] "South Carolina Mustard"
## [8] "dry rub"
## [9] "korean astyle"
## [10] "western style"
## [11] "other"
## [12] "Eastern Style (with no tomato)"
## [13] "Korean style"
## [14] "Eastern Style (no tomato)"
## [15] "Western Style (with tomato)"
## [16] NA
## [17] "Dry rub"
## [18] ". Kansas style (with molasses)"
## [19] "Western style (with no tomato)"
## [20] "Kansas Style (with molasses)"
## [21] "Western Style (with no tomato)"
## [22] "Eastern style (no tomato)"
## [23] "Kanas style (with molasses)"
## [24] "South Carolina mustard"
## [25] "Eastern style (with no tomato sauce)"
## [26] "Eastern Style (w/ no tomato)"
## [27] "Kansas Style (w/ no tomato)"
## [28] "Western Style (w/ no tomato)"
## [29] "Eastern style(with no tomato)"
## [30] "Western style(With tomato)"
## [31] "Eastern styel (with no tomato)"
## [32] "Kansas style (with malasses)"
```

```
df <- df %>%
  mutate(FavoriteSauce =
    case_match(
      FavoriteSauce,
      c("Eastern style (with no tomato)", "Eastern Style (with no tomato)",
        "Eastern Style (no tomato)", "Eastern style (no tomato)",
        "Eastern style (with no tomato sauce)", "Eastern Style (w/ no tomato)",
        "Eastern style(with no tomato)", "Eastern stye (with no tomato)") ~ "1",
      c("Western style (with tomato)", "western style", "Western Style (with tomato)",
        "Western style (with no tomato)", "Western Style (w/ no tomato)",
        "Western style(With tomato)", "Western Style (with no tomato)") ~ "2",
      c("Kansas style (with molasses)", ". Kansas style (with molasses)",
        "Kansas Style (with molasses)", "Kanas style (with molasses)",
        "Kansas Style (w/ no tomato)", "Kansas style (with malasses)") ~ "3",
      c("Dry Rub", "dry rub", "Dry rub") ~ "4",
      c("South Carolina Mustard", "South Carolina mustard") ~ "5",
      c("Korean Style", "korean astyle", "Korean style") ~ "6",
      c("Other", "other") ~ "7",
      .default = FavoriteSauce))
df$FavoriteSauce <- as.numeric(as.character(df$FavoriteSauce))
```

```
unique(df$FavoriteSide)
```

## Favorite Side

```
## [1] "Collard greens" "Hush puppies" "Baked beans" "Other"
## [5] "Fries" "Baked Beans" "Coleslaw" "Hush Puppies"
## [9] "Fried okra" "fries" "cheese grits" "collard greens"
## [13] "hush puppies" "baked beans" "Cheese grits" "Collard Greens"
## [17] "Fried Okra" "Green beans" "Cheese Grits" "Green Beans"
## [21] "Hushpuppies" "Okra" "Mac and cheese" "other"
## [25] "Colesaw" "Others" NA "Friend okra"
## [29] "fried okra" "coleslaw"
```

```
df <- df %>%
  mutate(FavoriteSide =
    case_match(
      FavoriteSide,
      c("Baked beans", "Baked Beans", "baked beans") ~ "1",
      c("Fried okra", "Fried Okra", "Okra", "Friend okra", "fried okra") ~ "2",
      c("Coleslaw", "Colesaw", "coleslaw") ~ "3",
      c("Hush puppies", "Hush Puppies", "hush puppies", "Hushpuppies") ~ "4",
      c("Green beans", "Green Beans") ~ "5",
      c("Fries", "fries") ~ "6",
      c("Collard greens", "collard greens", "Collard Greens") ~ "7",
      c("Cheese grits", "cheese grits", "Cheese Grits") ~ "8",
      c("Other", "other", "Others", "Mac and cheese") ~ "9",
      .default = FavoriteSide))
df$FavoriteSide <- as.numeric(as.character(df$FavoriteSide))
```

## Cleaning Numeric Variables

**Preferred Sauce Sweetness** Ensure all responses are numeric and remove responses outside of the specified scale (1-5):

```
df$Sweetness <- as.numeric(as.character(df$Sweetness))
df$Sweetness[df$Sweetness<1 | df$Sweetness>5] <- NA
```

```
df$Age <- as.numeric(as.character(df$Age))
df$Age[df$Age<5 | df$Age>100] <- NA
```

Age

```
df$MinutesDriving <- as.numeric(as.character(df$MinutesDriving))
df$MinutesDriving[df$MinutesDriving<0 | df$MinutesDriving>120] <- NA
```

**How Many Minutes Respondents Would Drive for BBQ**

```
df$SandwichPrice <- as.numeric(as.character(df$SandwichPrice))
df$SandwichPrice[df$SandwichPrice<0 | df$SandwichPrice>200] <- NA
```

**How Much Respondents Would Pay for a Pulled Pork Sandwich**

```
df$DinnerPlatePrice <- as.numeric(as.character(df$DinnerPlatePrice))
df$SandwichPrice[df$DinnerPlatePrice<0 | df$DinnerPlatePrice>200] <- NA
```

**How Much Respondents Would Pay for a Dinner Plate of Pulled Pork**

```
df$RibsPrice <- as.numeric(as.character(df$RibsPrice))
df$RibsPrice[df$RibsPrice<0 | df$RibsPrice>200] <- NA
```

**How Much Respondents Would Pay for Ribs** Check the dataset:

```
head(df, 20)
```

##	Observation	Sex	Age	Hometown	FavoriteMeat	FavoriteSauce	Sweetness
## 1	1	2	21	4	2	4	4
## 2	2	2	18	2	2	7	4
## 3	3	1	20	7	2	4	3
## 4	4	2	21	4	4	7	5
## 5	5	1	22	1	4	1	3
## 6	6	1	44	7	5	6	3
## 7	7	2	22	2	2	1	3
## 8	8	2	20	2	5	2	3
## 9	9	2	21	1	1	1	2
## 10	10	2	21	4	1	6	4
## 11	11	1	21	2	1	1	3
## 12	12	1	25	1	5	4	4
## 13	13	2	43	7	5	3	3
## 14	14	2	22	2	4	1	2
## 15	15	2	23	1	1	1	2
## 16	16	1	46	1	1	1	1
## 17	17	1	23	1	4	2	5
## 18	18	1	21	2	2	3	3
## 19	19	1	22	1	1	5	2
## 20	20	2	17	2	5	4	1

##	FavoriteSide	MinutesDriving	SandwichPrice	DinnerPlatePrice	RibsPrice
## 1	7	10	20	22	24
## 2	4	20	12	25	25
## 3	7	60	10	20	25
## 4	1	35	13	16	15
## 5	9	20	12	16	18
## 6	9	60	10	15	18
## 7	1	40	15	25	35
## 8	6	20	16	22	22
## 9	7	15	14	16	22
## 10	9	20	12	15	19
## 11	9	45	10	15	25
## 12	9	30	15	18	25
## 13	1	45	12	18	24
## 14	1	60	15	20	20
## 15	1	60	15	23	27
## 16	7	120	20	40	30
## 17	6	60	15	25	25
## 18	9	30	16	18	26
## 19	3	90	25	35	35
## 20	6	30	15	20	30

## Part Two: Data Visualization

### Bar Graphs

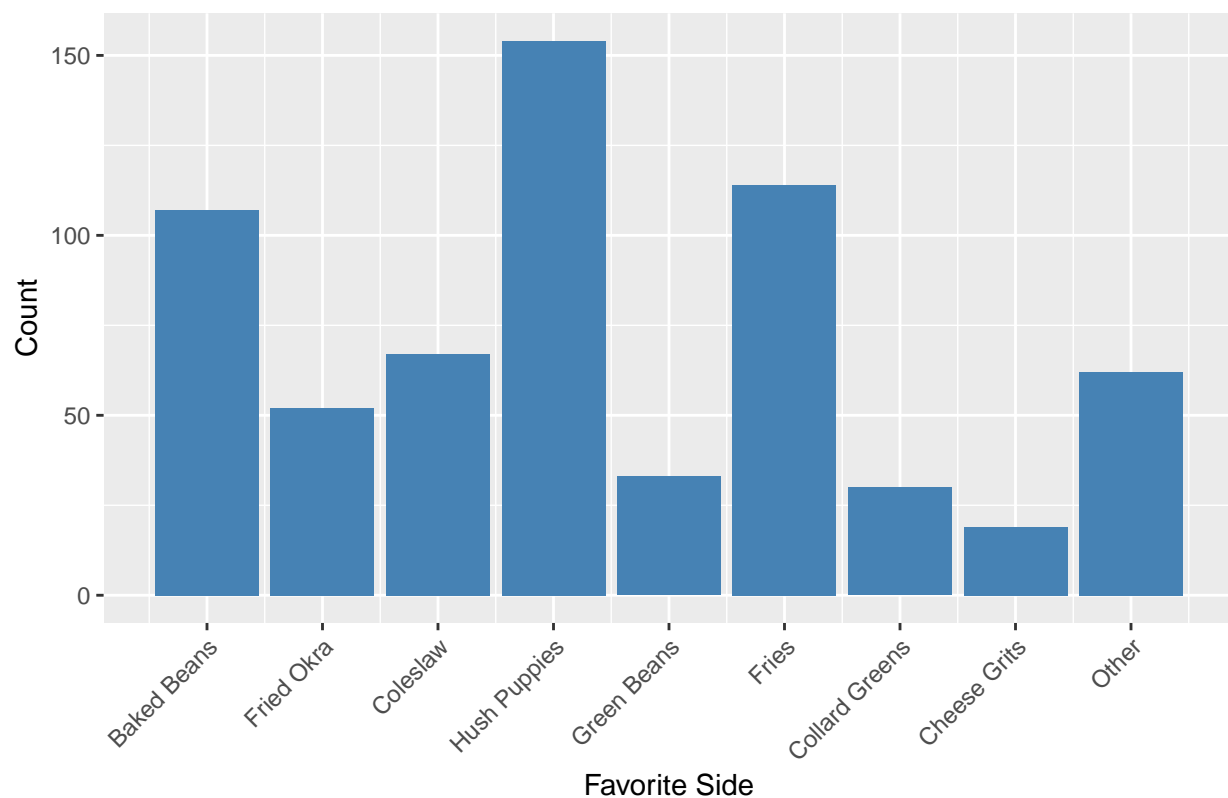
```
df %>%
  ggplot(aes(x = FavoriteSide)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Figure 1: Bar Graph of Favorite Sides",
```

```
x = "Favorite Side",
y = "Count") +
scale_x_continuous(breaks = c(1:9),
                    labels = c("Baked Beans", "Fried Okra", "Coleslaw",
                               "Hush Puppies", "Green Beans", "Fries",
                               "Collard Greens", "Cheese Grits", "Other")) +
theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
```

## Favorite Sides

## Warning: Removed 10 rows containing non-finite values ('stat\_count()').

Figure 1: Bar Graph of Favorite Sides



**Favorite Meat (Younger vs. Older Men)** Compare men 25 or younger to men older than 25 years old:

```
Men <- subset(df, df$Sex == 1)
Men <- Men %>%
  filter(!is.na(Age)) %>%
  mutate(AgeGroup = case_when(Age > 25 ~ "Older than 25",
                              Age <= 25 ~ "25 or younger"))
```

```
Men %>%
  ggplot(aes(x = FavoriteMeat, fill = AgeGroup)) +
  geom_bar(position = "dodge", stat = "count") +
```

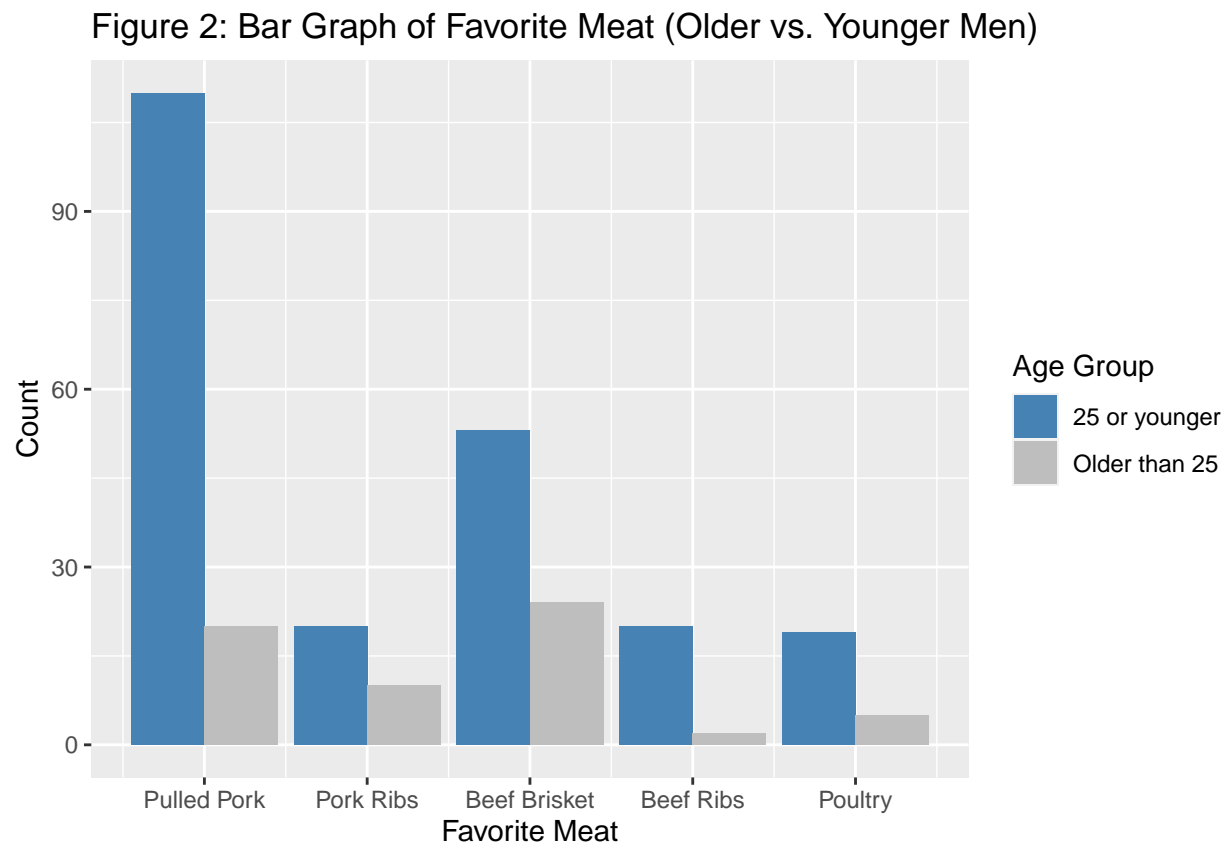


```

labs(title = "Figure 2: Bar Graph of Favorite Meat (Older vs. Younger Men)",
     x = "Favorite Meat",
     y = "Count",
     fill = "Age Group") +
scale_x_continuous(breaks = c(1:5),
                  labels = c("Pulled Pork", "Pork Ribs", "Beef Brisket",
                             "Beef Ribs", "Poultry")) +
scale_fill_manual(values = c("steelblue", "gray"))

```

## Warning: Removed 1 rows containing non-finite values ('stat\_count()').



## Histograms

```

df %>%
  ggplot(aes(x = Age)) +
  geom_histogram(binwidth = 3,
                color = "black", fill = "steelblue") +
  scale_x_continuous(breaks=seq(11, 92, 3), lim = c(11, 92)) +
  labs(title = "Figure 3: Histogram of Age",
       x = "Age",
       y = "Count")

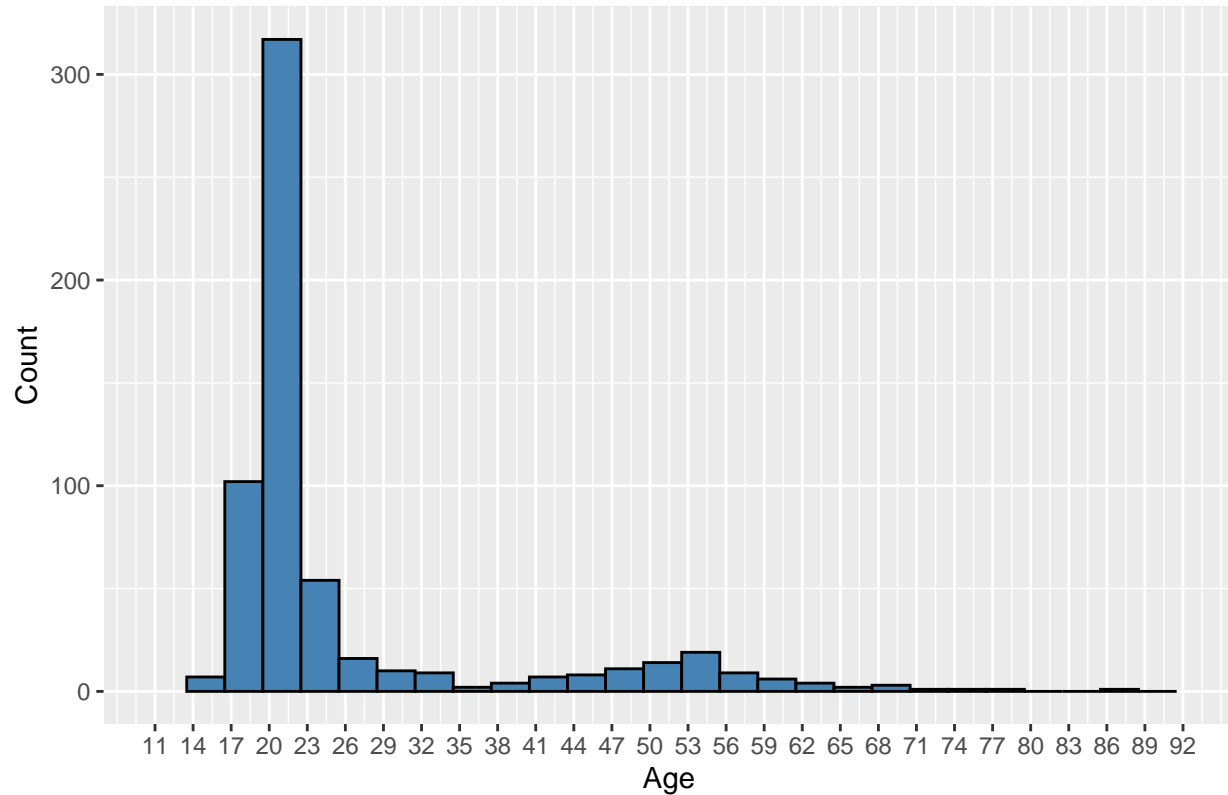
```

## Age

## Warning: Removed 39 rows containing non-finite values ('stat\_bin()').

## Warning: Removed 2 rows containing missing values ('geom\_bar()').

Figure 3: Histogram of Age



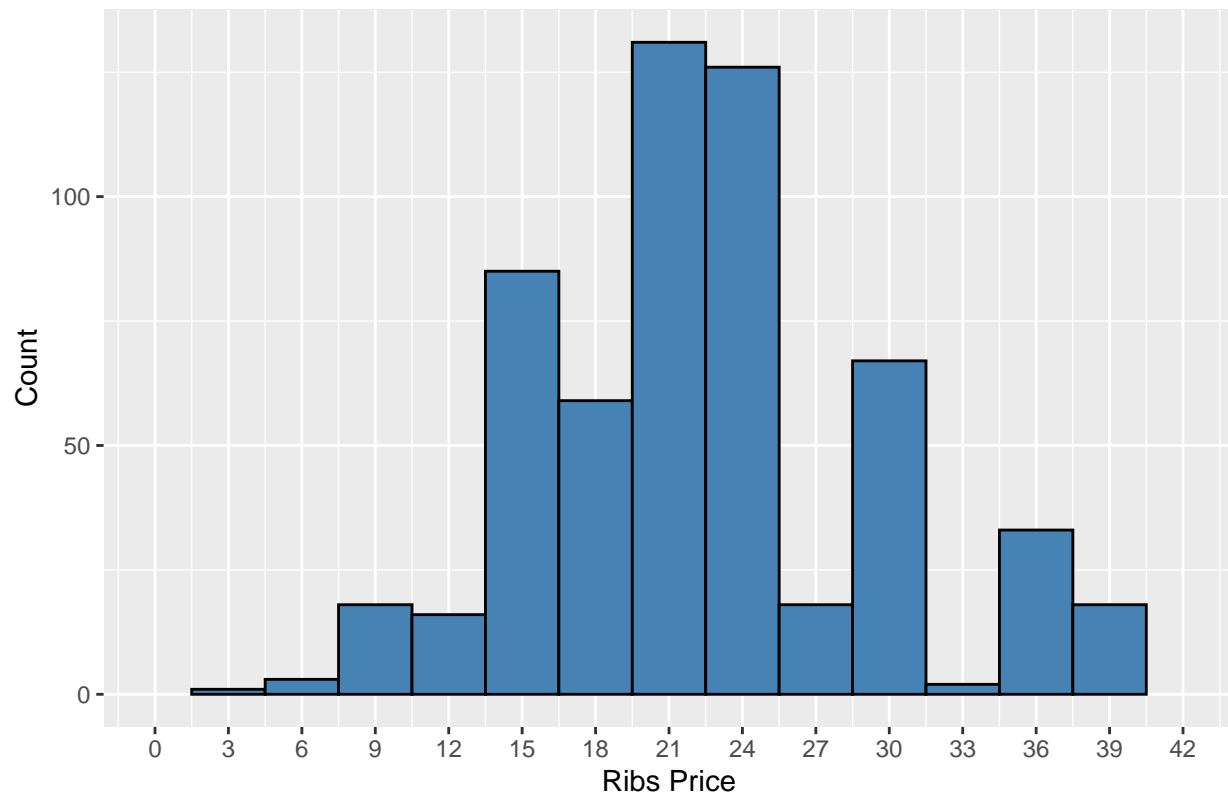
```
df %>%  
  ggplot(aes(x = RibsPrice)) +  
  geom_histogram(binwidth = 3,  
                 color = "black", fill = "steelblue") +  
  scale_x_continuous(breaks=seq(0, 42, 3), lim = c(0, 42)) +  
  labs(title = "Figure 4: Histogram of Ribs Price",  
       x = "Ribs Price",  
       y = "Count")
```

## How Much Respondents Would Pay for Ribs

## Warning: Removed 45 rows containing non-finite values ('stat\_bin()').

## Warning: Removed 2 rows containing missing values ('geom\_bar()').

Figure 4: Histogram of Ribs Price



## Scatterplots

**Age and Minutes Driving** Remove outliers (130 and above):

```
df %>%  
  filter(MinutesDriving < 130) %>%  
  ggplot(aes(x = Age, y = MinutesDriving)) +  
  geom_point() +  
  geom_smooth(method = lm, se = FALSE) +  
  labs(title = "Figure 5: Scatterplot of Age and Minutes Driving (Outliers Removed)",  
       x = "Age",  
       y = "Minutes Driving")
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```

```
## Warning: Removed 34 rows containing non-finite values ('stat_smooth()').
```

```
## Warning: Removed 34 rows containing missing values ('geom_point()').
```

Figure 5: Scatterplot of Age and Minutes Driving (Outliers Removed)

