Problem Set 1 Key

1. The data set `Arrests` can be accessed through the package `car` in`R`. The data set includes data on police treatment of individuals arrested in Toronto for simple possession of small quantities of marijuana. The data are part of a larger data set featured in a series of articles in the Toronto Star newspaper.

- Provide the relevant `R` code to install the `car` package in `R`.

```
install.packages("car")
```

- Provide the relevant `R` code to load the `car` package.

```
library(car)
```

```
Loading required package: carData
```

- Explore the `Arrests` data set using `glimpse()` to show the different variables available in the data set. What are the different variables in the `Arrests` data set? Show your code. [4pts.]

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.5.1
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()    masks stats::lag()
x dplyr::recode() masks car::recode()
x purrr::some()   masks car::some()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
glimpse(Arrests)
```

```
Rows: 5,226
Columns: 8
$ released <fct> Yes, No, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes~
$ colour   <fct> White, Black, White, Black, Black, Black, White, White, Black~
$ year     <int> 2002, 1999, 2000, 2000, 1999, 1998, 1999, 1998, 2000, 2001, 1~
$ age      <int> 21, 17, 24, 46, 27, 16, 40, 34, 23, 30, 18, 18, 17, 42, 26, 2~
$ sex      <fct> Male, Male, Male, Male, Female, Female, Male, Female, Male, M~
$ employed <fct> Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Ye~
$ citizen  <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Ye~
$ checks   <int> 3, 3, 3, 1, 1, 0, 0, 1, 4, 3, 0, 3, 1, 0, 2, 3, 4, 5, 3, 0, 0~
```

- Create a frequency table for the year of arrests (year variable). Which year had the most arrests? Show your code. [2pts.]

```
library(summarytools)
```

```
Warning: package 'summarytools' was built under R version 4.5.1
```

```

Attaching package: 'summarytools'

The following object is masked from 'package:tibble':

    view
```

```
freq(Arrests$year)
```

```
Error in match(x, table, nomatch = 0L): 'match' requires vector arguments

Warning in parse_call(mc = match.call(), caller = "freq"): metadata extraction
terminated unexpectedly; inspect results carefully
```

Frequencies

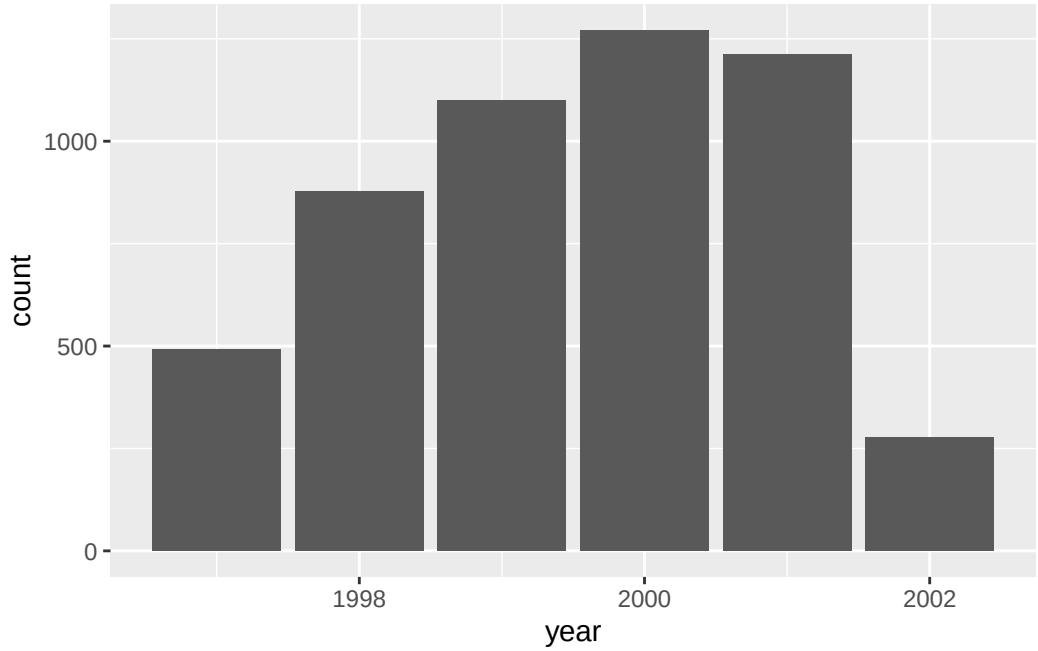
	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
1997	492	9.41	9.41	9.41	9.41
1998	877	16.78	26.20	16.78	26.20
1999	1099	21.03	47.23	21.03	47.23
2000	1270	24.30	71.53	24.30	71.53
2001	1211	23.17	94.70	23.17	94.70
2002	277	5.30	100.00	5.30	100.00
<NA>	0			0.00	100.00
Total	5226	100.00	100.00	100.00	100.00

> **Note**
>
> The year with the most arrests was 2000.

- Create a bar plot to visualize the frequencies of arrests for each year using ggplot() and geom_bar(). Show your code. [2pts.]

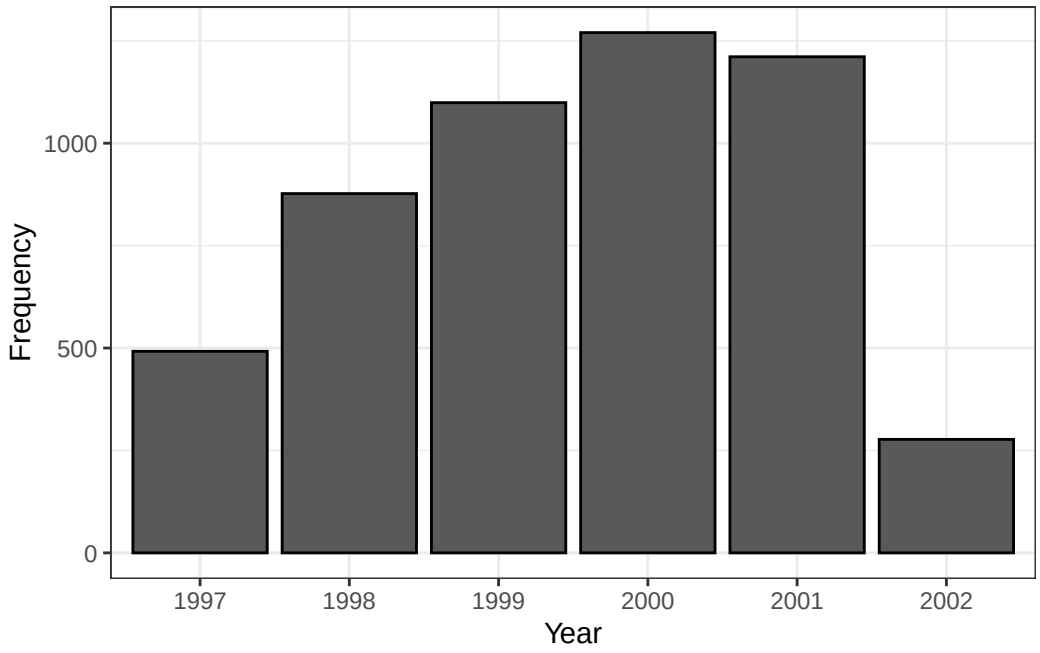
Acceptable solution (full credit):

```
ggplot(data=Arrests,aes(x=year)) +
  geom_bar()
```



Better solution:

```
ggplot(data=Arrests,aes(x=as.factor(year))) +
  geom_bar(color="black") +
  theme_bw() +
  labs(x="Year", y="Frequency")
```



2. The file `SleepHealthData.csv` comprises 400 rows and 13 columns, covering a wide range of variables related to sleep and daily habits. It includes details such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders.

- Provide the relevant `R` code to import `SleepHealthData.csv` into `R` and assign it to the variable `sleep`.

> **!** Important
>
> Save the CSV file in the same folder as your `R`script!

```
getwd() # provides your working directory
```

```
sleep <- read.csv("SleepHealthData.csv")
```

- Complete the following table for the recorded heart rate of the participants. Show your code.

Mean:

```
mean(sleep$heart_rate)
```

```
[1] 70.16578
```

Median:

```
median(sleep$heart_rate)
```

```
[1] 70
```

Range:

```
max(sleep$heart_rate)-min(sleep$heart_rate)
```

```
[1] 21
```

Variance and Standard Deviation:

```
var(sleep$heart_rate)
```

```
[1] 17.10381
```

```
sd(sleep$heart_rate)
```

```
[1] 4.135676
```

Minimum and Maximum:

```
min(sleep$heart_rate)
```

```
[1] 65
```

```
max(sleep$heart_rate)
```

```
[1] 86
```

Q_1, Q_3, and IQR:

```
quantile(sleep$heart_rate,0.25)
```

```
25%
 68
```

```
quantile(sleep$heart_rate,0.75)
```

```
75%
 72
```

```
IQR(sleep$heart_rate)
```

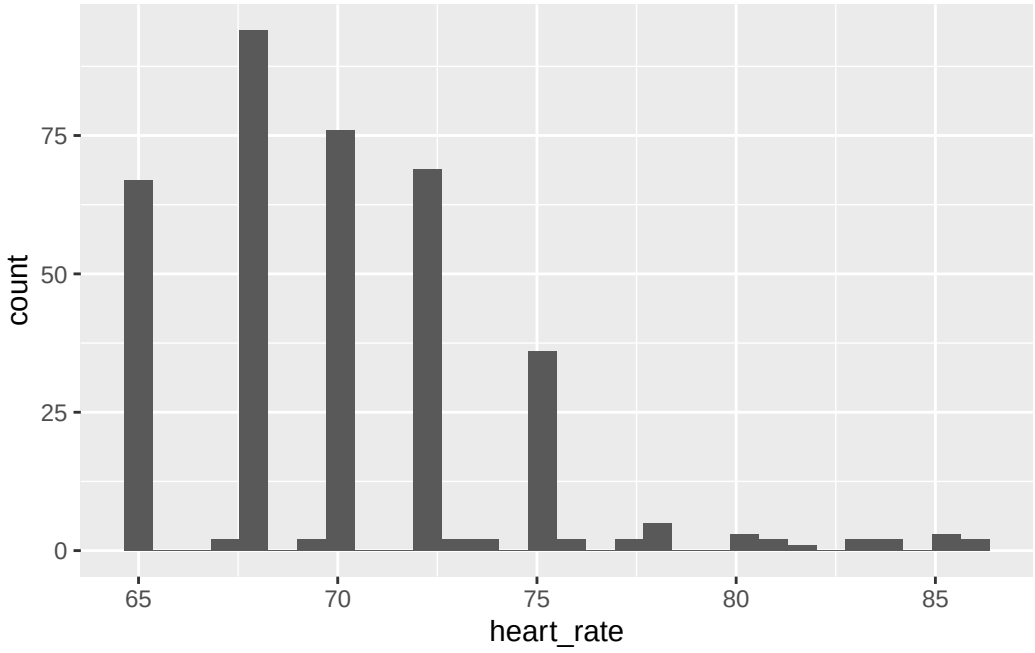
```
[1] 4
```

- Create a histogram for the recorded heart rate of all the participants using `ggplot()` and `geom_histogram()`. Show your code.

Acceptable (full credit):

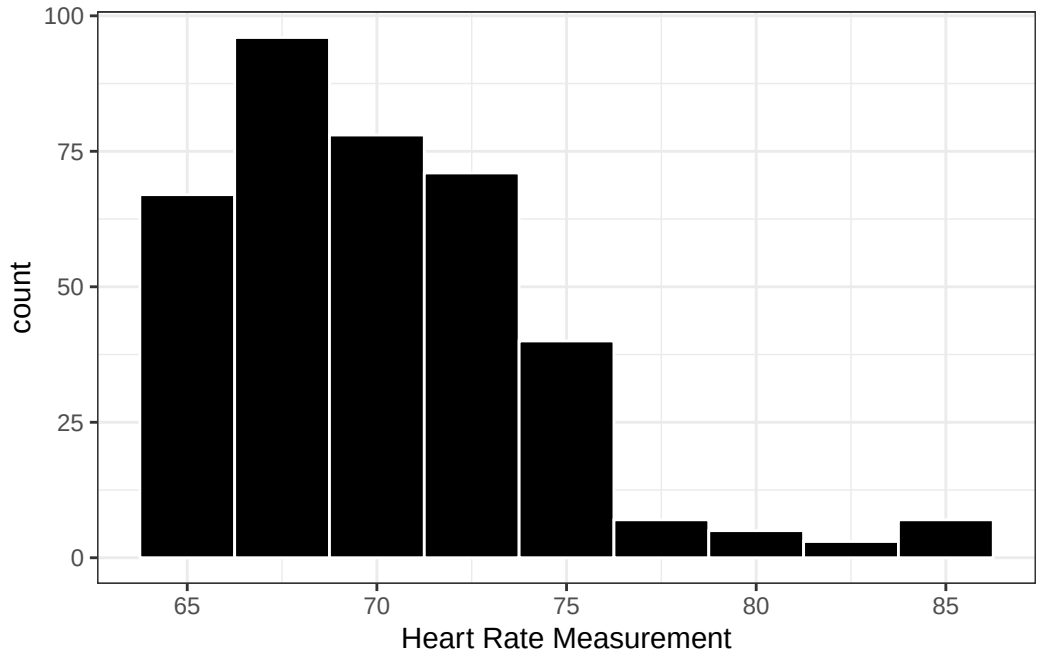
```
ggplot(data=sleep, aes(x=heart_rate)) +
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



Better version:

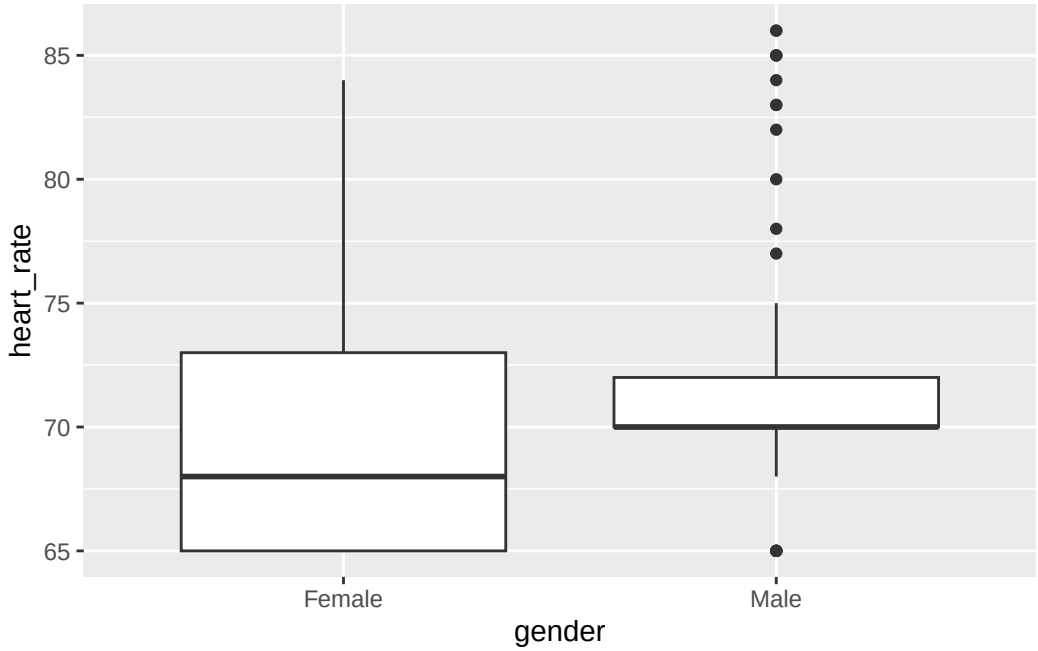
```
ggplot(data=sleep, aes(x=heart_rate)) +
  geom_histogram(color="white", binwidth=2.5, fill="black") +
  theme_bw() +
  labs(x="Heart Rate Measurement")
```



- Create a boxplot for the recorded heart rate of all the participants grouped by gender using ggplot() and geom_boxplot(). Show your code. (Hint: Use x=gender in the aes() function.)

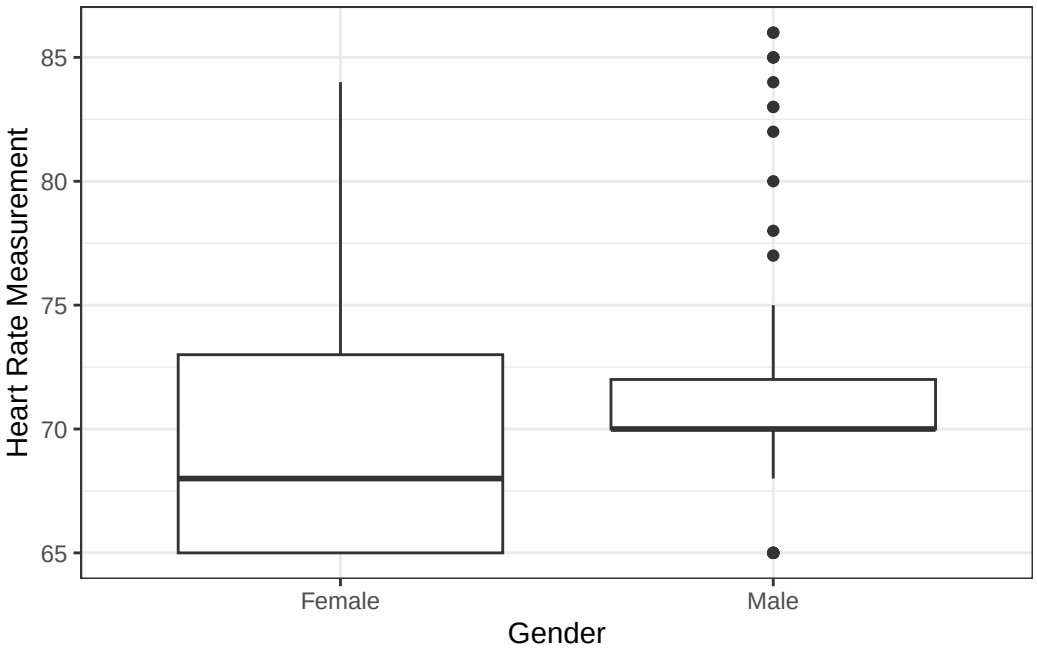
Acceptable (full credit):

```
ggplot(data=sleep, aes(x=gender,y=heart_rate)) +
  geom_boxplot()
```



Better version:

```
ggplot(data=sleep, aes(x=gender, y=heart_rate)) +
  geom_boxplot() +
  theme_bw() +
  labs(x="Gender",y="Heart Rate Measurement")
```


Heart Rate Measurement
85
80
75
70
65
Female
Male
Gender