Homework 14

Alizeh Azhar

12/1/2021

Dataset Selection

The dataset that I have chosen for my final project is the *Housing dataset*, which contains *Sales Prices of Houses in the City of Windsor*. The dataset contains 546 number of observations and 12 variables. Out of the 12 variables, 6 are numeric and 6 are categorical. Table 1 below shows the summary of the dataframe.

```
Housing <- read.csv("C:/Users/hp/Downloads/Housing.csv")
Housing= Housing[,-1]

library(tibble)
Housing<- tibble(Housing)
# str(Housing)
library(skimr)
skim_without_charts(Housing)
```

Table 1: Data summary

Name	Housing
Number of rows	546
Number of columns	12
Column type frequency:	
character	6
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
driveway	0	1	2	3	0	2	0
recroom	0	1	2	3	0	2	0
fullbase	0	1	2	3	0	2	0
gashw	0	1	2	3	0	2	0
airco	0	1	2	3	0	2	0
prefarea	0	1	2	3	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
price	0	1	68121.60	26702.67	25000	49125	62000	82000	190000
lotsize	0	1	5150.27	2168.16	1650	3600	4600	6360	16200
bedrooms	0	1	2.97	0.74	1	2	3	3	6
bathrms	0	1	1.29	0.50	1	1	1	2	4
stories	0	1	1.81	0.87	1	1	2	2	4
garagepl	0	1	0.69	0.86	0	0	0	1	3

Exploratory Analysis and Data Summary

In this section, all the 12 variables in the dataset are summarized.

Price

It is a numeric continuous variable that represents the sale price of a house. In our analysis, it will be the dependent variable. The table below shows the summary statistics of the variable.

```
library(summarytools)
```

```
##
## Attaching package: 'summarytools'
```

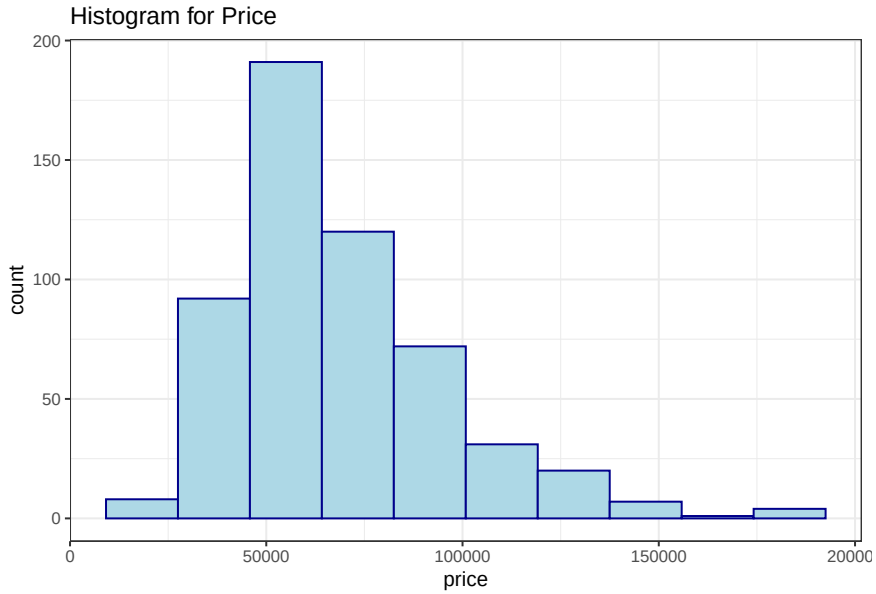
```
## The following object is masked from 'package:tibble':
##
##     view
```

```
descr(Housing$price, stats = c("mean", "sd", "min", "q1", "med", "q3", "max"), transpose = T, display.la
```

	Mean	Std.Dev	Min	Q1	Median	Q3	Max
price	68121.60	26702.67	25000.00	49000.00	62000.00	82000.00	190000.00

We can summarize the distribution of the `price` variable using a histogram.

```
library(ggplot2)
ggplot(data= Housing, aes(x= price))+
  geom_histogram(bins = 10, color="darkblue",fill= "lightblue")+
  theme_bw()+
  ggtitle("Histogram for Price")
```



From the histogram above, the peak of the price variable occurs where the average sale prices of he houses is 50,000. The data is right skewed. Most of the sample values are clustered on the right side of the histogram. The number of houses decreases with increase in price. Most houses that were observed had low prices.

Lot size

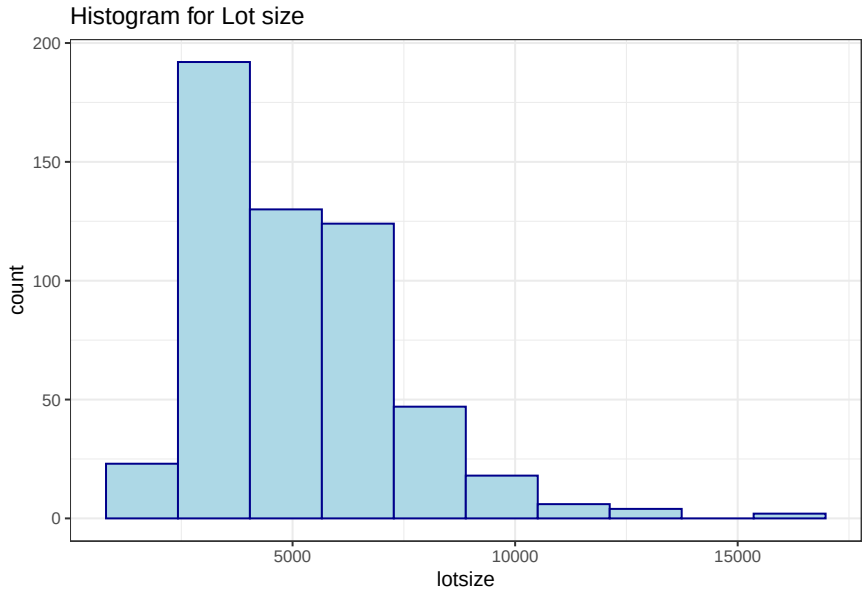
It is a numeric variable that represents the lot size of a house in square feet. In our analysis, it will be one of the independent variables. The table below shows the summary statistics of the variable.

```
descr(Housing$lotsize, stats = c("mean", "sd", "min", "q1", "med", "q3", "max"), transpose = T, display
```

	Mean	Std.Dev	Min	Q1	Median	Q3	Max
lotsize	5150.27	2168.16	1650.00	3600.00	4600.00	6360.00	16200.00

We can summarize the variable graphically using an histogram

```
ggplot(data= Housing, aes(x= lotsize))+
  geom_histogram(bins = 10, color="darkblue",fill= "lightblue")+
  theme_bw()+
  ggtitle("Histogram for Lot size")
```



The lot size of the houses has a mean of 5150.27 and the standard deviation of 2168.16. The lot size has a right skew but has only one peak. From the position of the peak, number of houses are decreasing as the lot size is increasing. From the histogram, there are some visible outliers in the lot size.

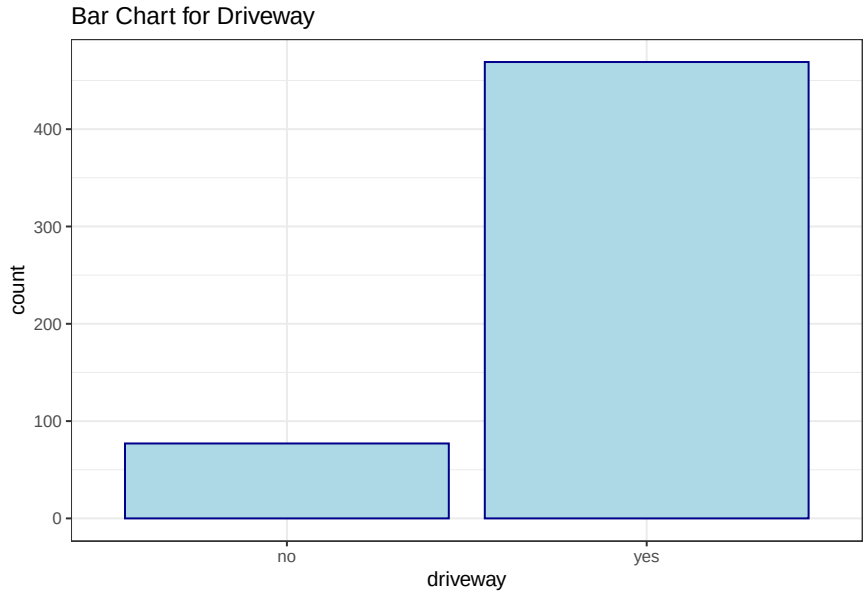
Driveway

This is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a driveway, otherwise, it is *no*. In this analysis, it will be an independent variable. Below is the frequency table and the bar chart of the driveway variable.

```
library(summarytools)
freq(Housing$driveway, report.nas = F, headings = F)
```

	Freq	%	% Cum.
no	77	14.10	14.10
yes	469	85.90	100.00
Total	546	100.00	100.00

```
library(ggplot2)
ggplot(data = Housing, aes(driveway, fill=driveway))+
  geom_bar(color="darkblue",fill= "lightblue")+
  scale_fill_manual(values = c("#E69F00","#56B4E9"))+
  theme_bw()+
  ggtitle("Bar Chart for Driveway")
```



From the bar plot above, the number of houses that had a driveway were approximately 6 times the number of houses that did not have a driveway.

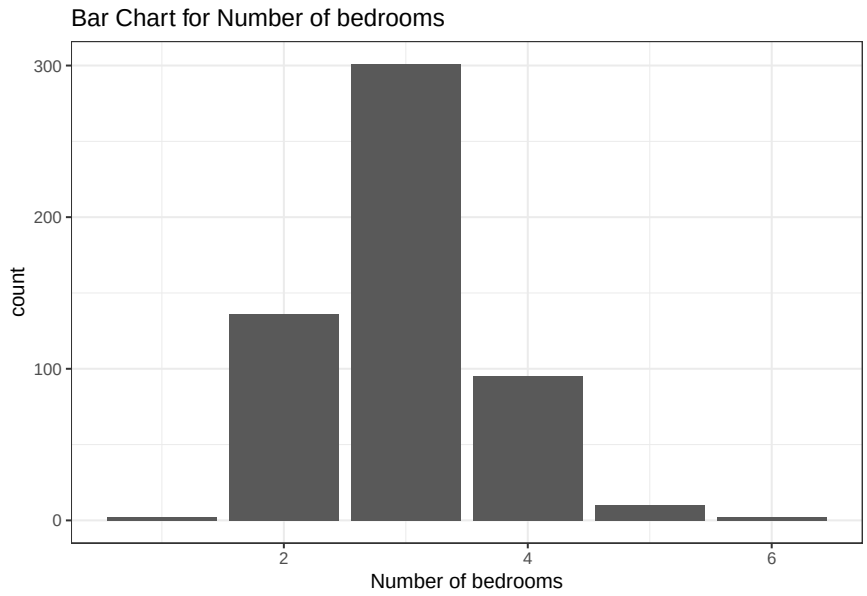
Number of bedrooms

The number of bedrooms is represented by *bedrooms*. It is an independent discrete variable. However, it will not be used in this analysis. Below is the summary statistics and bar chart of the number of bedrooms.

```
freq(Housing$bedrooms, report.nas = F, headings = F)
```

	Freq	%	% Cum.
1	2	0.37	0.37
2	136	24.91	25.27
3	301	55.13	80.40
4	95	17.40	97.80
5	10	1.83	99.63
6	2	0.37	100.00
Total	546	100.00	100.00

```
library(ggplot2)
ggplot(data = Housing, aes(x=bedrooms))+
  geom_bar()+
  theme_bw()+
  xlab("Number of bedrooms")+
  ggtitle("Bar Chart for Number of bedrooms")
```



The number of bedrooms range from 1 to 6. A few houses had less than 2 bedrooms. Most houses had 2 and 3 bedrooms, and the number of bedrooms increased.

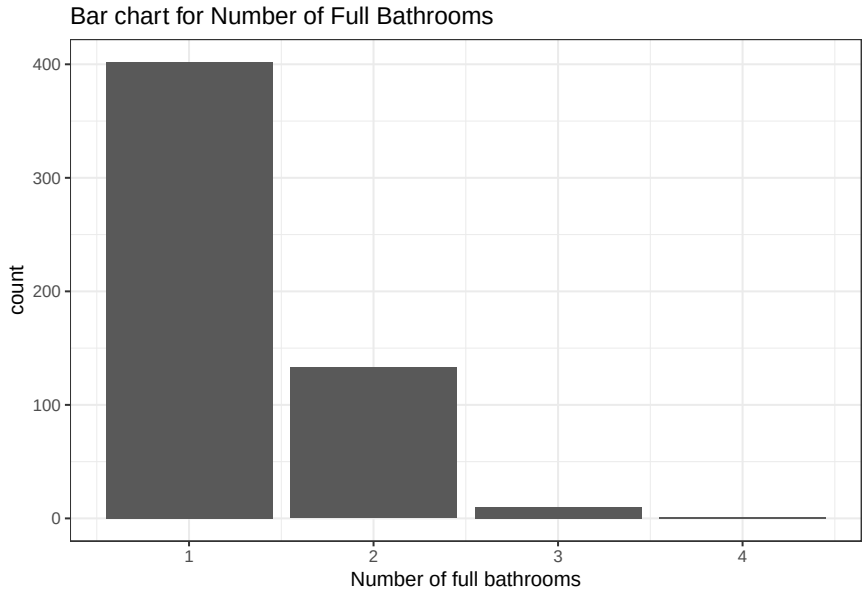
Number of full bathrooms

The number of full bathrooms is represented by *bathrms*. It is an independent discrete variable. However, it will not be used in this analysis, despite that it has an effect on the prices of houses. Below is the table of summary statistics and bar chart of the number of bathrooms.

```
freq(Housing$bathrms, report.nas = F, headings = F)
```

	Freq	%	% Cum.
1	402	73.63	73.63
2	133	24.36	97.99
3	10	1.83	99.82
4	1	0.18	100.00
Total	546	100.00	100.00

```
ggplot(data= Housing, aes(x= bathrms))+
  geom_bar()+
  theme_bw()+
  xlab("Number of full bathrooms")+
  ggtitle("Bar chart for Number of Full Bathrooms")
```



The number of bathrooms range from 1 to 4. Most houses had one bathroom the least number of houses had 4 bedrooms.

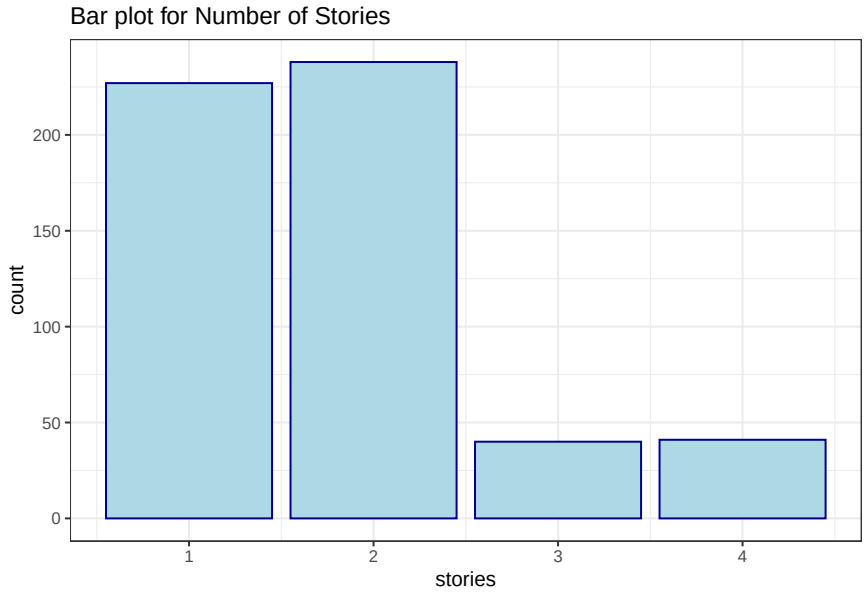
Stories

This variable contains the number of stories excluding basement. It is a discrete independent variable. This variable will not be used in our analysis. Below is the summary statistics and the bar chart of the *stories* variable.

```
freq(Housing$stories, report.nas = F, headings = F)
```

	Freq	%	% Cum.
1	227	41.58	41.58
2	238	43.59	85.16
3	40	7.33	92.49
4	41	7.51	100.00
Total	546	100.00	100.00

```
library(ggplot2)
ggplot(data= Housing, aes(x= stories))+
  geom_bar(color="darkblue",fill= "lightblue")+
  theme_bw()+
  ggtitle("Bar plot for Number of Stories")
```



The number of stories range from 1 to 4. Most houses had 2 stories. The number of houses that had 3 and 4 stories have almost equal frequencies.

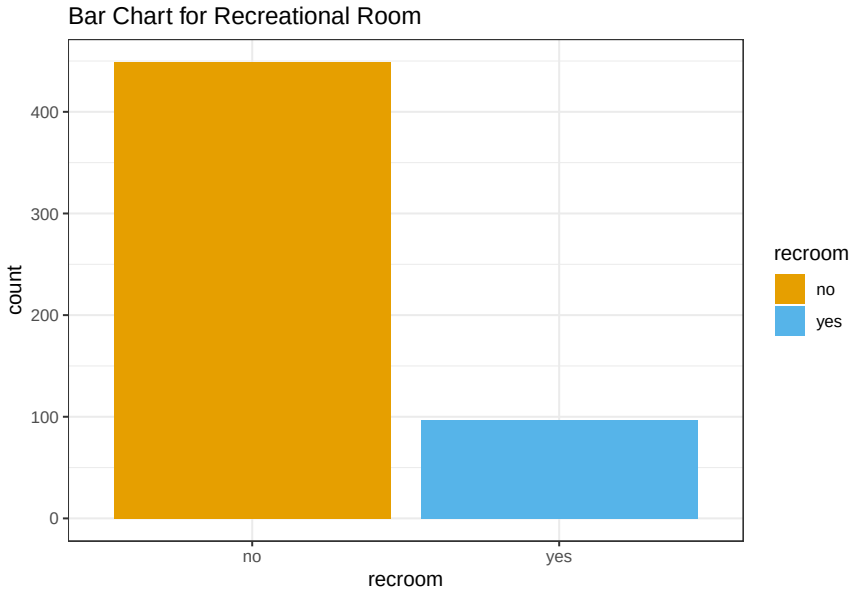
Recreational room

In the dataset, the variable is represented by `recroom`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a recreational room, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

```
library(summarytools)
freq(Housing$recroom, report.nas = F, headings = F)
```

	Freq	%	% Cum.
no	449	82.23	82.23
yes	97	17.77	100.00
Total	546	100.00	100.00

```
library(ggplot2)
ggplot(data = Housing, aes(recroom, fill=recroom))+
  geom_bar()+
  scale_fill_manual(values = c("#E69F00","#56B4E9"))+
  theme_bw()+
  ggtitle("Bar Chart for Recreational Room")
```



Most houses did not have a recreation room. Only a few houses had a recreation room.

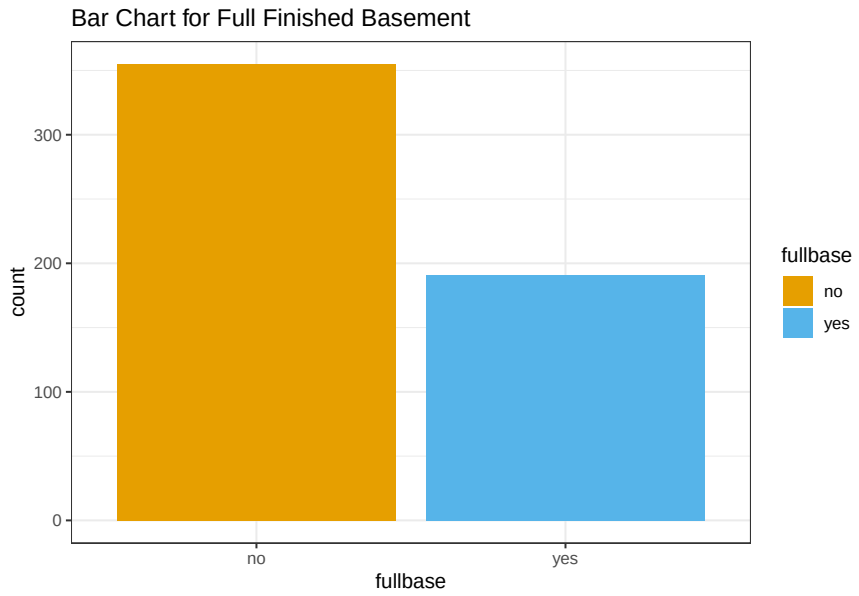
Full finished basement

In the dataset, the variable is represented by `fullbase`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a full finished basement, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

```
library(summarytools)
freq(Housing$fullbase, report.nas = F, headings = F)
```

	Freq	%	% Cum.
no	355	65.02	65.02
yes	191	34.98	100.00
Total	546	100.00	100.00

```
library(ggplot2)
ggplot(data = Housing, aes(fullbase, fill=fullbase))+
  geom_bar()+
  scale_fill_manual(values = c("#E69F00","#56B4E9"))+
  theme_bw()+
  ggtitle("Bar Chart for Full Finished Basement")
```



From the above bar chart, most houses had a furnished basement.

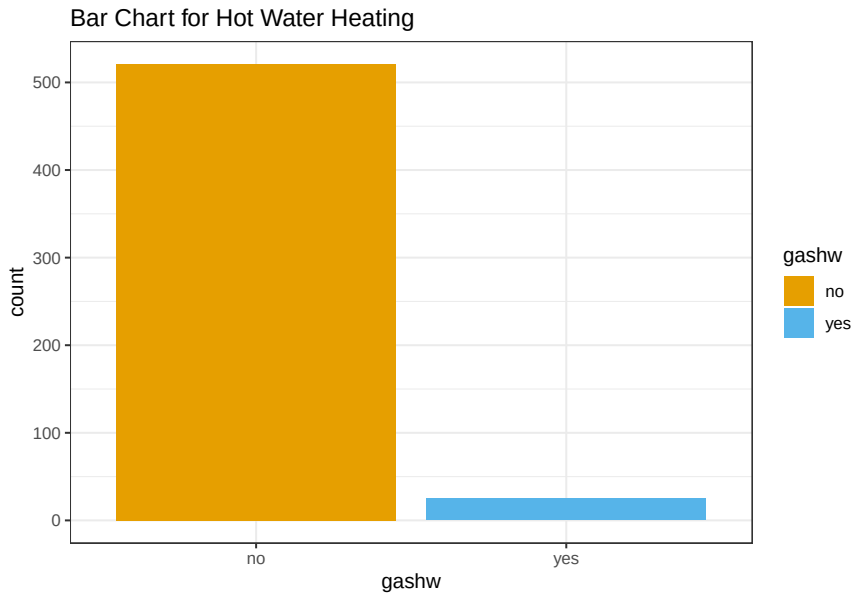
Gas for hot water heating

In the dataset, the variable is represented by `gashw`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a gas for hot water heating, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

```
library(summarytools)
freq(Housing$gashw, report.nas = F, headings = F)
```

	Freq	%	% Cum.
no	521	95.42	95.42
yes	25	4.58	100.00
Total	546	100.00	100.00

```
library(ggplot2)
ggplot(data = Housing, aes(gashw, fill=gashw))+
  geom_bar()+
  scale_fill_manual(values = c("#E69F00","#56B4E9"))+
  theme_bw()+
  ggtitle("Bar Chart for Hot Water Heating")
```



Most houses had a gas for heating water. Very few houses did not have a gas for heating water

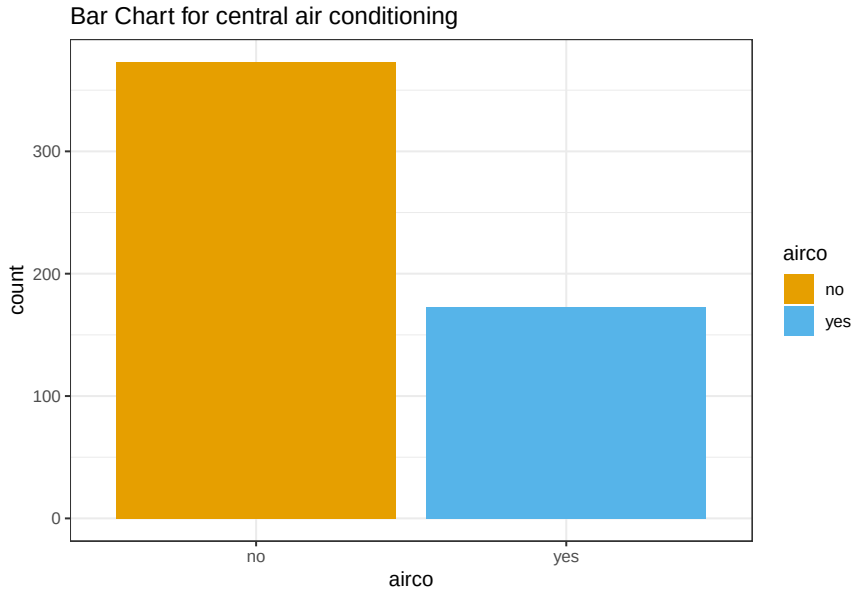
Central air conditioning

In the dataset, the variable is represented by `airco`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a central air conditioning, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

```
library(summarytools)
freq(Housing$airco, report.nas = F, headings = F)
```

	Freq	%	% Cum.
no	373	68.32	68.32
yes	173	31.68	100.00
Total	546	100.00	100.00

```
library(ggplot2)
ggplot(data = Housing, aes(airco, fill=airco))+
  geom_bar()+
  scale_fill_manual(values = c("#E69F00","#56B4E9"))+
  theme_bw()+
  ggtitle("Bar Chart for central air conditioning")
```



More than half of the observed houses had a central air conditioning.

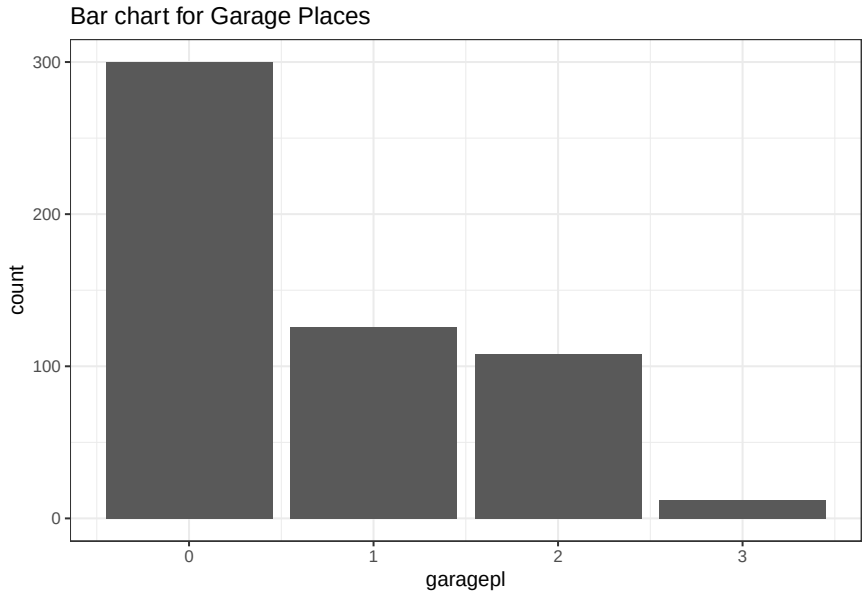
Garage places

The number of garage places is represented by *garagepl*. It is an independent descrete variable. However, it will not be used in this analysis. Below is the summary statistics and histogram of the number of garage places.

```
library(summarytools)
freq(Housing$garagepl, report.nas = F, headings = F)
```

	Freq	%	% Cum.
0	300	54.95	54.95
1	126	23.08	78.02
2	108	19.78	97.80
3	12	2.20	100.00
Total	546	100.00	100.00

```
ggplot(data= Housing, aes(garagepl, fill=garagepl))+
  geom_bar()+
  scale_fill_manual(values = c("#E69F00","#56B4E9"))+
  theme_bw()+
  ggtitle("Bar chart for Garage Places")
```



The number of garage places range from 0 to 3. Most houses did not have a garage place. The frequency decreased as the number of garage places increased.

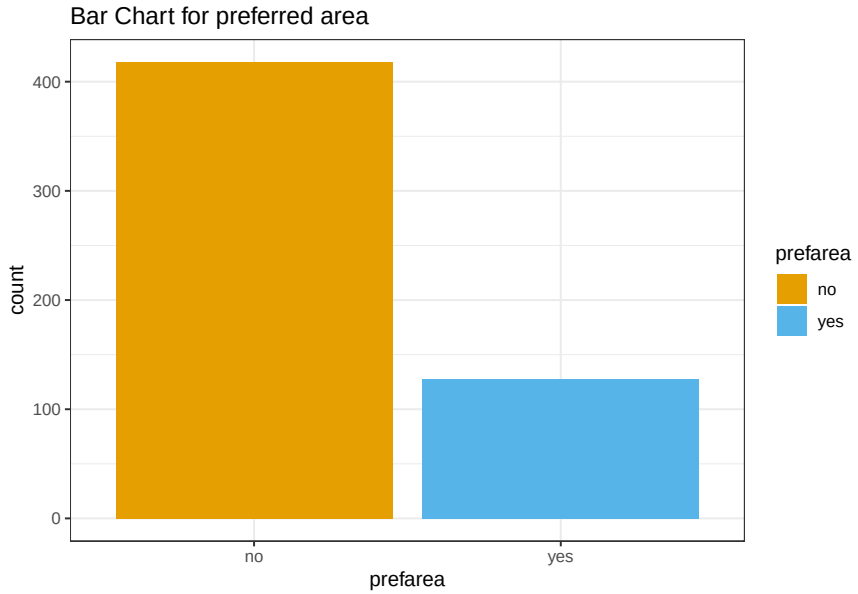
Preferred area

In the dataset, the variable is represented by `prefarea`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house located in the preferred neighborhood of the city, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

```
freq(Housing$airco, report.nas = F, headings = F)
```

	Freq	%	% Cum.
no	373	68.32	68.32
yes	173	31.68	100.00
Total	546	100.00	100.00

```
library(ggplot2)
ggplot(data = Housing, aes(prefarea, fill=prefarea))+
  geom_bar()+
  scale_fill_manual(values = c("#E69F00","#56B4E9"))+
  theme_bw()+
  ggtitle("Bar Chart for preferred area")
```



More than half of the observed houses were not located in the preferred neighborhood of the city.