STA141 Midterm Exam 2

Richard McCormick

2023-04-21

INSTRUCTIONS: You may use any materials available, including your notes, textbook, and online information. If any information is used, details must be given on what was used and how it works.

Please prepare your solutions using the RMD file provided. You may change options to suit your style, but be sure to keep the document organized. *Justify all free response answers. Type solutions after each prompt and try to keep your PDF organized.*

Please submit your exam as an organized PDF document directly from RMD. The solutions should be presented in the order the questions were asked. If there are problems with your PDF you will be asked to resubmit. Organization, clarity and correct preparation of solutions will be worth **5 points**.

This exam is worth a total of 100 points, which is 20 percent of you overall grade for this course.

Question 1 (40 points)

The file `Sacramento.csv` contains house and sale price data for 932 homes in Sacramento CA.

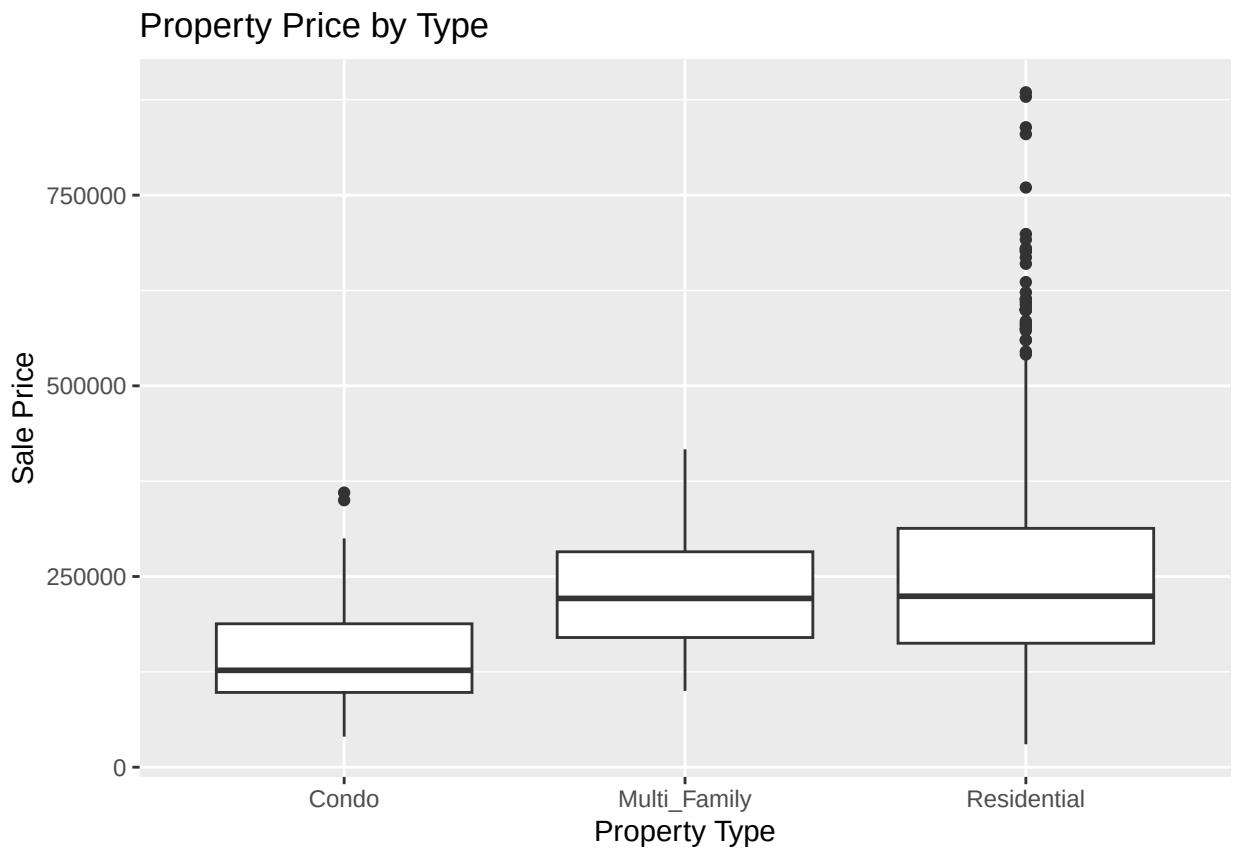
(a) Import the data from the csv file provided.

```
sacramento.data <- read_csv( 'Sacramento.csv' )
```

```
## New names:
## Rows: 932 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (3): city, zip, type dbl (7): ...1, beds, baths, sqft, price, latitude,
## longitude
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

(b) Create a boxplot showing the distribution of the sale price against each type of home (there should be multiple boxes on the one plot).

```
ggplot( data=sacramento.data, aes( x=type, y=price ) ) +
  geom_boxplot() +
  labs( title="Property Price by Type", x="Property Type", y="Sale Price" )
```



(c) How much cheaper is the median Condo price than the median Multi-family home price?

```
median.price <- sacramento.data %>%
  group_by( type ) %>%
  summarize( median( price ) )

median.price
```

```
## # A tibble: 3 x 2
##   type         `median(price)`
##   <chr>                  <dbl>
## 1 Condo                127000
## 2 Multi_Family         221250
## 3 Residential          224126
```

```
print( "The median Condo price is $94,250 cheaper than the median Multi-Family home." )
```

```
## [1] "The median Condo price is $94,250 cheaper than the median Multi-Family home."
```

(d) The boxplot for Condos appears to have outliers. What price is the upper threshold for an outlier in this case? How many condos were sold for a value that is greater than this?

```
condo.data <- sacramento.data[sacramento.data$type=="Condo",]
```

```
upper.thresh <- quantile( condo.data$price )[4] + (1.5 * IQR( condo.data$price ))

print( paste( "The upper threshold for outliers for Condo price is: $", upper.thresh ) )
```

```
## [1] "The upper threshold for outliers for Condo price is: $ 323000"
```

```
print( paste( "The number of Condos sold for MORE than the upper theshold is:",
              sum( condo.data$price > upper.thresh ) ) )
```

```
## [1] "The number of Condos sold for MORE than the upper theshold is: 2"
```

(e) The boxplot for residential homes also has outliers. What is the difference between the mean and median price of a residential home?

```
median.price
```

```
## # A tibble: 3 x 2
##   type         `median(price)`
##   <chr>                  <dbl>
## 1 Condo                127000
## 2 Multi_Family         221250
## 3 Residential          224126
```

```
mean.price <- sacramento.data %>%
  group_by( type ) %>%
  summarize( mean( price ) )

mean.price
```

```
## # A tibble: 3 x 2
##   type         `mean(price)`
##   <chr>                <dbl>
## 1 Condo              148669.
## 2 Multi_Family       224535.
## 3 Residential        252991.
```

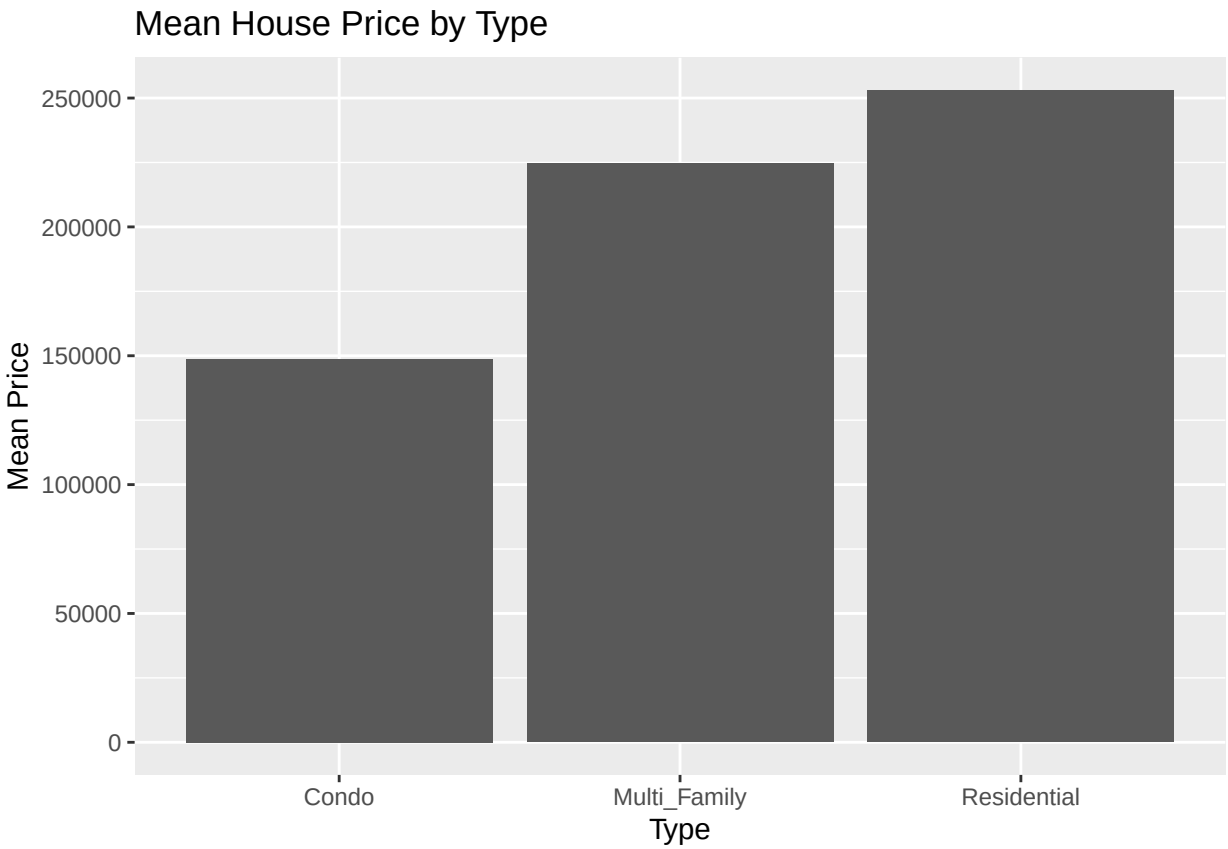
```
print( "The price difference between median and mean sales price for a residential home is $3,284.7")
```

```
## [1] "The price difference between median and mean sales price for a residential home is $3,284.7"
```

(f) Plot a bar chart showing the mean price of a home against the type of home.

```
mean.house <- sacramento.data %>%
  group_by( by=type ) %>%
  summarize( mean.price=mean( price ) )

ggplot( data=mean.house, aes( x=by, y=mean.price ) ) +
  geom_bar( stat='identity' ) +
  labs( title="Mean House Price by Type", x="Type", y="Mean Price" )
```



Question 2 (20 points)

The Systolic blood pressures of adults, in the appropriate units, are normally distributed with a mean of 128.4 and a standard deviation of 19.6. We are going to use this to simulate and analyze a population of adults.

The following code block with create a random sample to simulate a fictional city of 50,000 people. Run the following code block and then answer the questions below using this data.

```
my.population <- data_frame(bp=rnorm(50000, mean=128.4, sd=19.6))
```

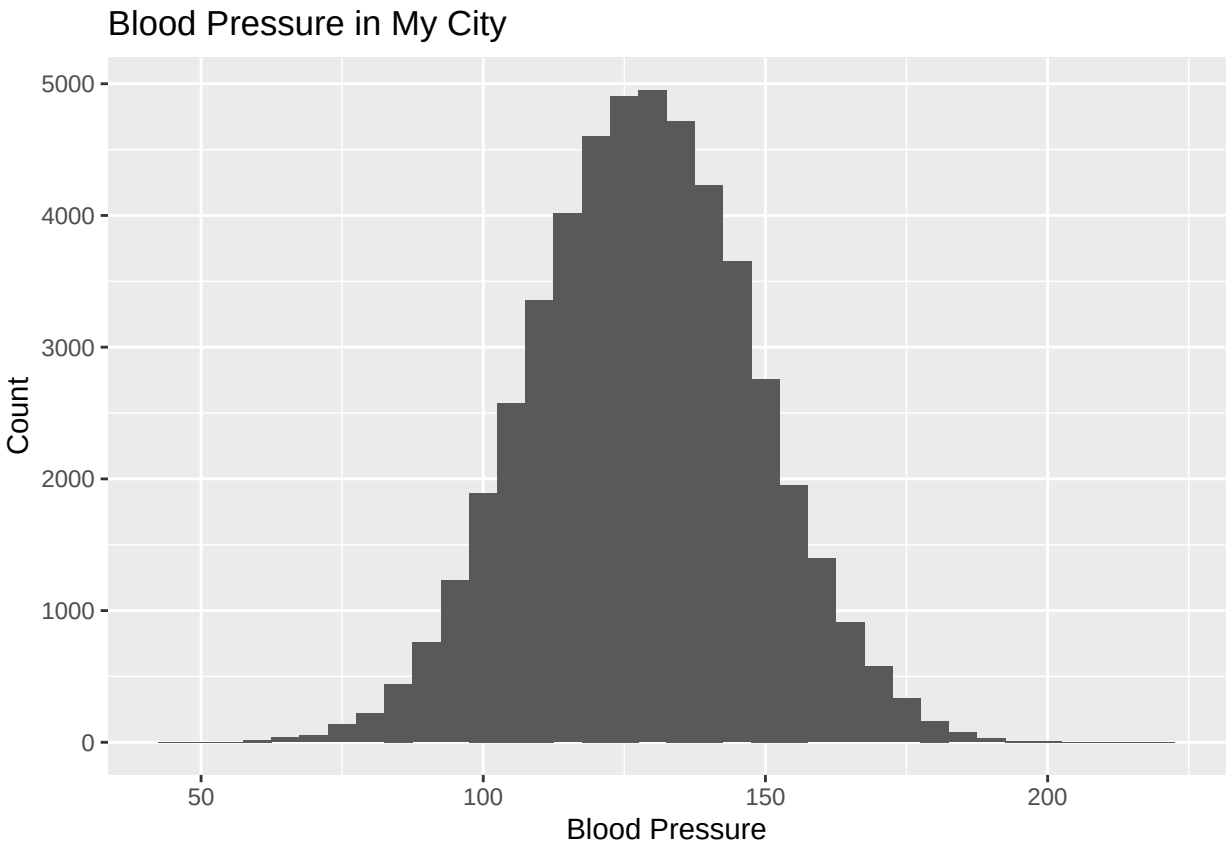
(a) Find the mean Blood Pressure of your city.

```
print( paste( "Mean blood pressue is:", mean( my.population$bp ) ) )
```

```
## [1] "Mean blood pressue is: 128.275875000028"
```

(b) Plot a histogram of the population's Blood Pressure using a binwidth of 5 units.

```
ggplot( data=my.population ) +
  geom_histogram( aes( x=bp ), binwidth=5 ) +
  labs( title="Blood Pressure in My City", x="Blood Pressure", y="Count" )
```



(c) High Blood Pressure is defined as having a Systolic Blood Pressure above 130. What percentage of your city's population has High Blood Pressure?

```
high.bp <- ( sum( my.population$bp > 130 ) / 50000 ) * 100

print( paste( "The proportion of my city's population with high blood pressure is:",
              high.bp, "%" ) )
```

```
## [1] "The proportion of my city's population with high blood pressure is: 46.562 %"
```

Question 3 (40 points)

The file `basketball.csv` shows the total points scored by selected NBA teams during the last week.

(a) Import the data from the csv file provided.

```
basketball.data <- read_csv( 'basketball.csv' )
```

```
## Rows: 4 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Team
## dbl (4): Game 1, Game 2, Game 3, Game 4
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

(b) This data is in the wide format. Use `pivot_longer` to convert it to the long format.

```
basketball.long <- pivot_longer( basketball.data,
                                 cols=2:5,
                                 values_to = "Points",
                                 names_to = "Game" )

basketball.long
```

```
## # A tibble: 16 x 3
##    Team     Game   Points
##    <chr>    <chr>   <dbl>
##  1 Phoenix  Game 1     88
##  2 Phoenix  Game 2    103
##  3 Phoenix  Game 3    105
##  4 Phoenix  Game 4     94
##  5 Chicago  Game 1     91
##  6 Chicago  Game 2    120
##  7 Chicago  Game 3     94
##  8 Chicago  Game 4     96
##  9 Brooklyn Game 1     89
## 10 Brooklyn Game 2    131
## 11 Brooklyn Game 3     99
## 12 Brooklyn Game 4    101
## 13 Houston  Game 1     94
## 14 Houston  Game 2     92
## 15 Houston  Game 3     92
## 16 Houston  Game 4     98
```

(c) Find the mean number of points for each team in this period.

```
basketball.long %>%
  group_by( Team ) %>%
  summarize( mean( Points ) )
```

```
## # A tibble: 4 x 2
##   Team     `mean(Points)`
##   <chr>             <dbl>
## 1 Brooklyn          105
## 2 Chicago           100.
## 3 Houston            94
## 4 Phoenix            97.5
```