Lab 1

AUTHOR
JP Garcia

Question 1: Part a

```
library(readr)
rent_data <- read_csv("~/jp-garcia-131a/craigslist.csv")
```

```
Rows: 5876 Columns: 7
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (3): title, link, location
dbl  (3): price, size, brs
dttm (1): time

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.
```

```
num_postings <- nrow(rent_data)
print(paste("The number of postings in the craiglist dataset is ", num_postings))
```

```
[1] "The number of postings in the craiglist dataset is  5876"
```

Question 1: Part b

```
#calculation
mean_rent <- mean(rent_data$price)
median_rent <- median(rent_data$price)
max_rent <- max(rent_data$price)
min_rent <- min(rent_data$price)

print(paste("Mean Monthly Rent:", mean_rent))
```

```
[1] "Mean Monthly Rent: 3125.22515316542"
```

```
print(paste("Median Monthly Rent:", median_rent))
```

```
[1] "Median Monthly Rent: 2865"
```

```
print(paste("Max Monthly Rent:", max_rent))
```

```
[1] "Max Monthly Rent: 20000"
```

```
print(paste("Min Monthly Rent :", min_rent))
```

```
[1] "Min Monthly Rent : 600"
```

Question 1: Part c

```
city_postings <- table(rent_data$location)

print("City Distribution:")
```

```
[1] "City Distribution:"
```

```
print(city_postings)
```

```
           alameda albany / el cerrito            berkeley          emeryville 
               200                111                396                210 
        menlo park       mountain view             oakland           palo alto 
               389               1052                881                710 
      redwood city            richmond          sunnyvale 
               521                451                955 
```

Question 1: Part d

```
total_postings <- nrow(rent_data)
postings_over_3000 <- sum(rent_data$price > 3000)
percent_over_3000 <- (postings_over_3000 / total_postings) * 100

print(paste("Entries Over 3000:", round(percent_over_3000, 2), "%"))
```

```
[1] "Entries Over 3000: 42.31 %"
```

Question 2

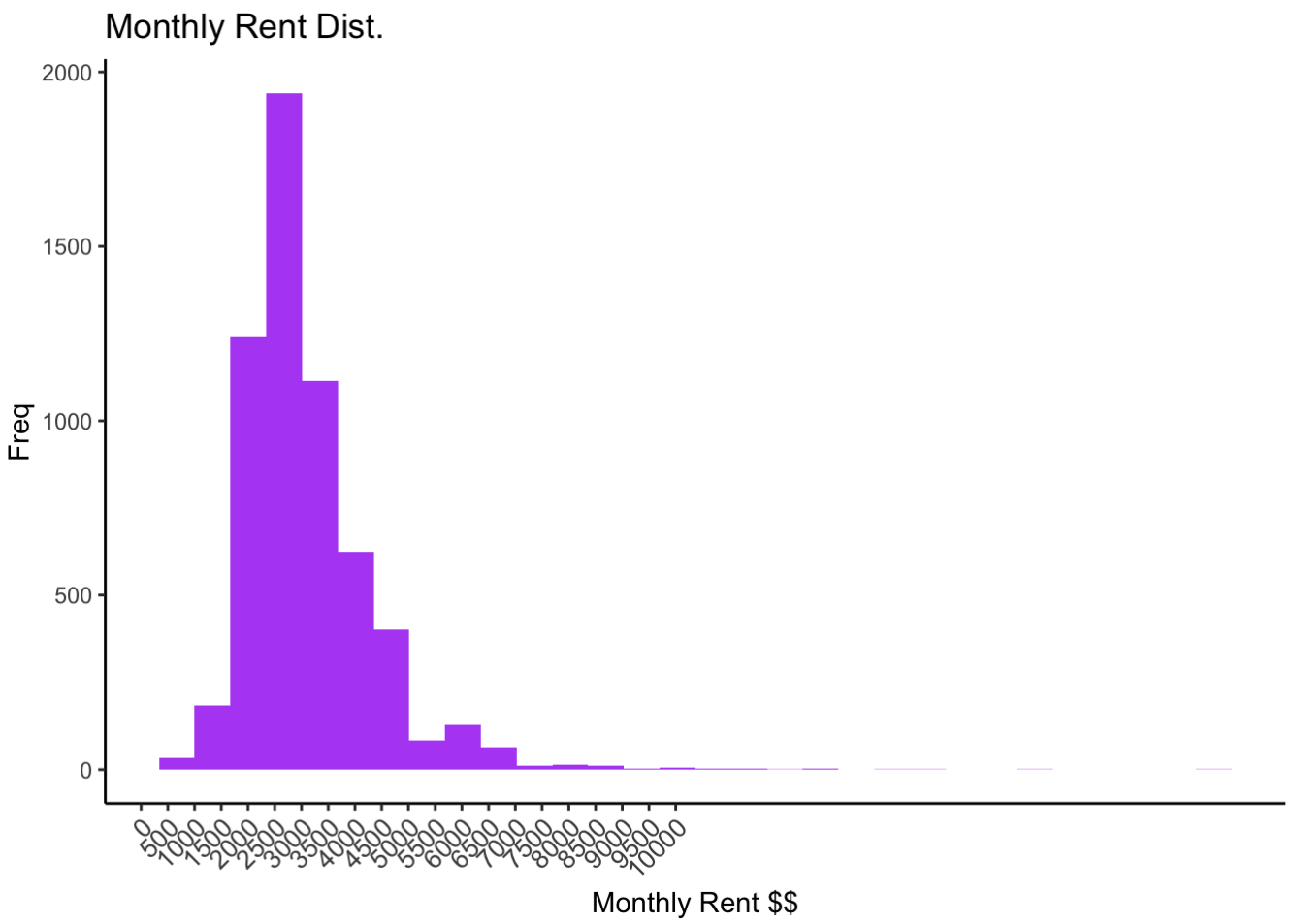
```
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
craigslist <- read.csv("~/jp-garcia-131a/craigslist.csv")

craigslist %>%
```

```
  ggplot(aes(x = price)) +
  geom_histogram(bins = 30, fill = "purple", alpha = 0.8) +
  lims(x = c(0, 10000)) +
  scale_x_continuous(breaks = seq(from = 0, to = 10000, by = 500)) +
  labs(
    title = "Monthly Rent Dist.",
    x = "Monthly Rent $$",
    y = "Freq"
  ) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10))
```

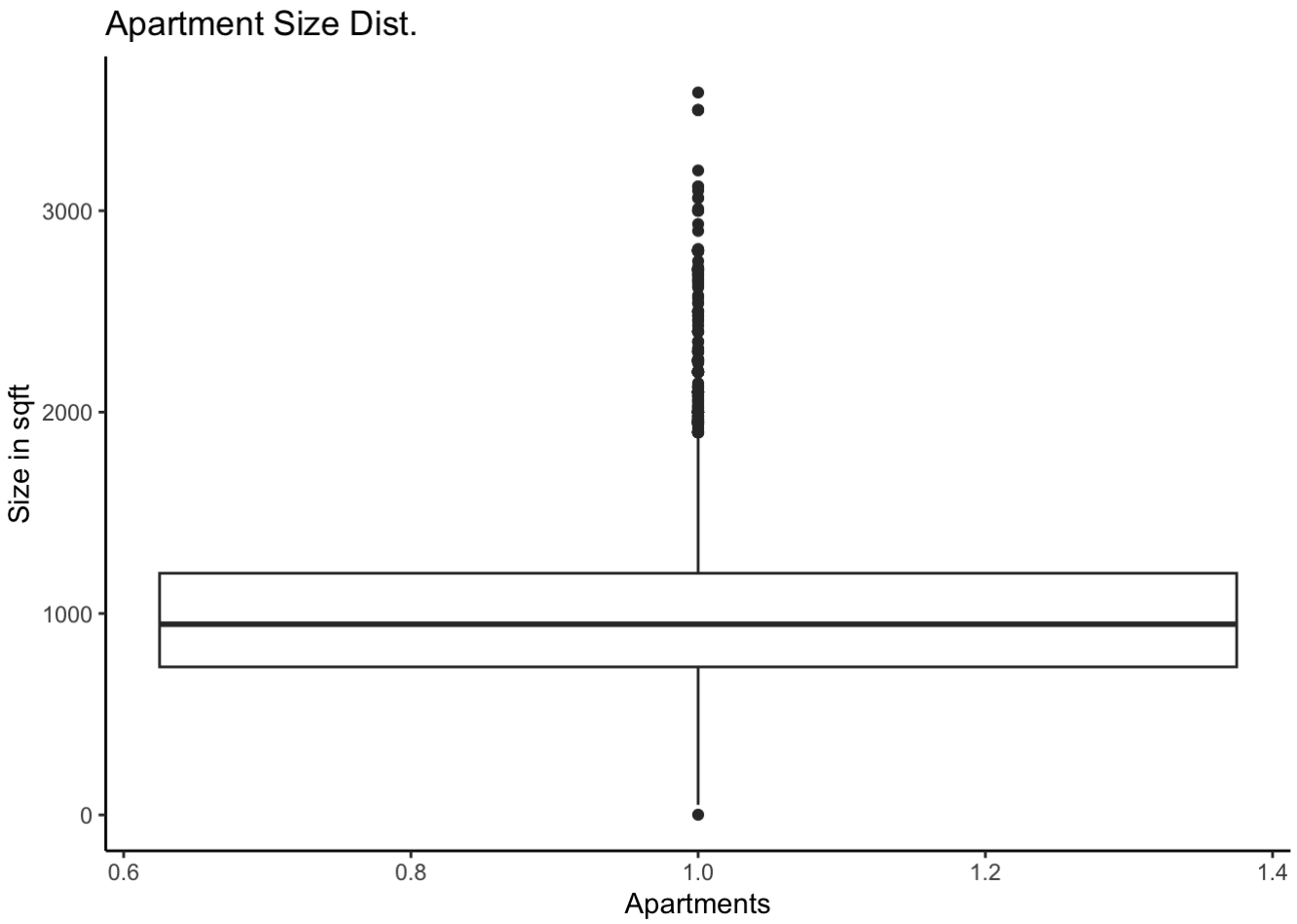
```
Scale for x is already present.
Adding another scale for x, which will replace the existing scale.
```



Question 3: Part a

```
library(ggplot2)
library(dplyr)
craigslist <- read.csv("~/jp-garcia-131a/craigslist.csv")

craigslist %>%
  filter(size <= 60000) %>%
  ggplot(aes(x = 1, y = size)) +
```

```
  geom_boxplot() +
  labs(
    title = "Apartment Size Dist. ",
    x = "Apartments",
    y = "Size in sqft"
  ) +
  theme_classic()
```

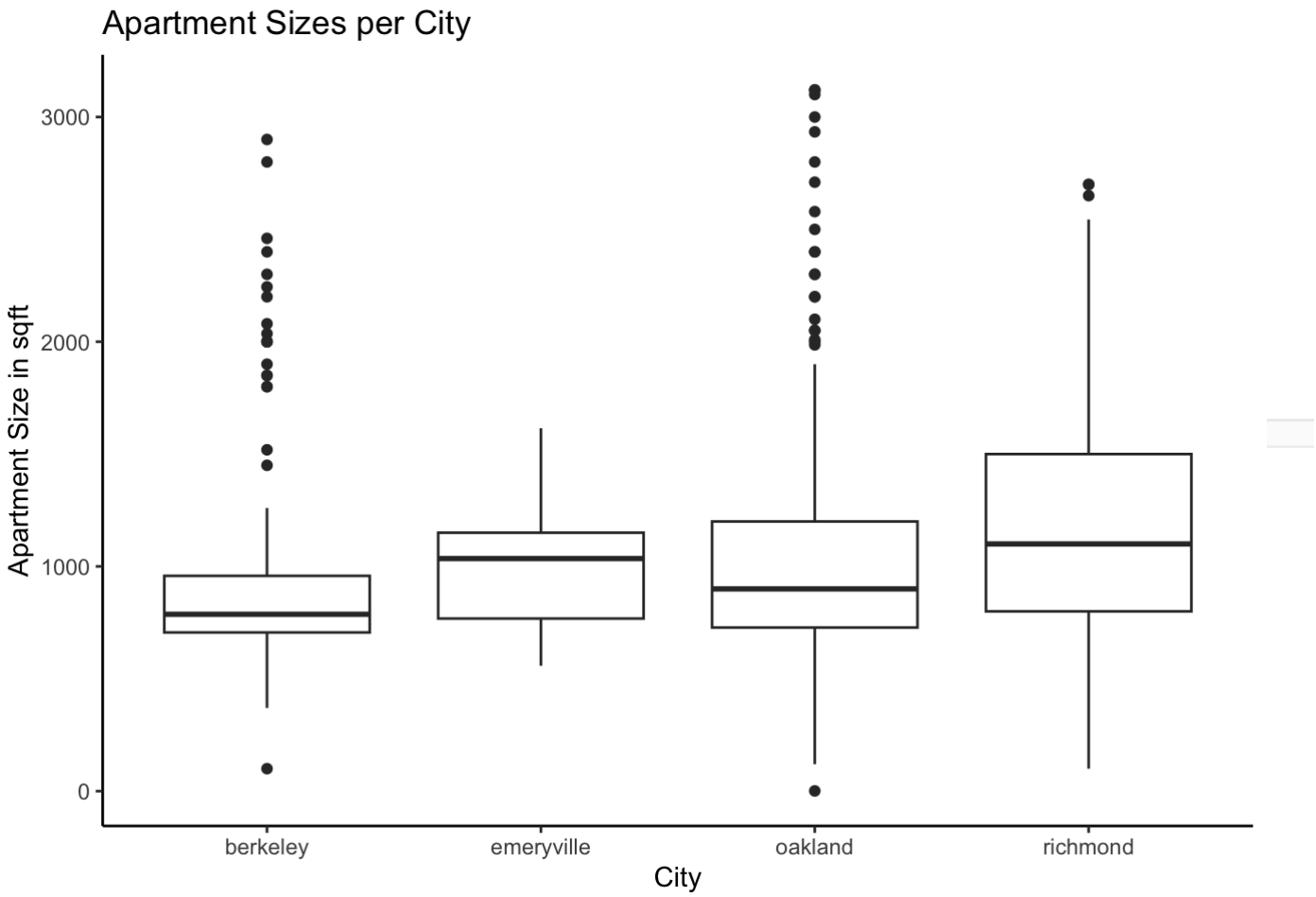


Question 3: Part b

```
library(ggplot2)
library(dplyr)

filtered_listings <- craigslist %>%
  filter(location %in% c("berkeley", "oakland", "richmond", "emeryville", "albany", "el c

ggplot(data = filtered_listings, aes(x = location, y = size)) +
  geom_boxplot() +
  labs(
    title = "Apartment Sizes per City",
    x = "City",
    y = "Apartment Size in sqft"
```

```
  ) +
  theme_classic()
```

```
Warning: Removed 761 rows containing non-finite values (`stat_boxplot()`).
```



Intro to Probability: Question 1

```
library(ggplot2)
library(dplyr)

median_rent<- median(craigslist$price, na.rm = TRUE)

print(median_rent)
```

```
[1] 2865
```

```
filtered_rent <- craigslist %>%
                 filter(price < median_rent)

ggplot(data = filtered_rent, aes(x = price)) +
  geom_histogram(binwidth = 100, fill = "grey", alpha = 0.75) +
  labs(
      title = "Distribution of Rent Prices Less Than Median",
```

```
      x = "Monthly Rent",
      y = "Freq"
  ) +
  theme_classic()
```

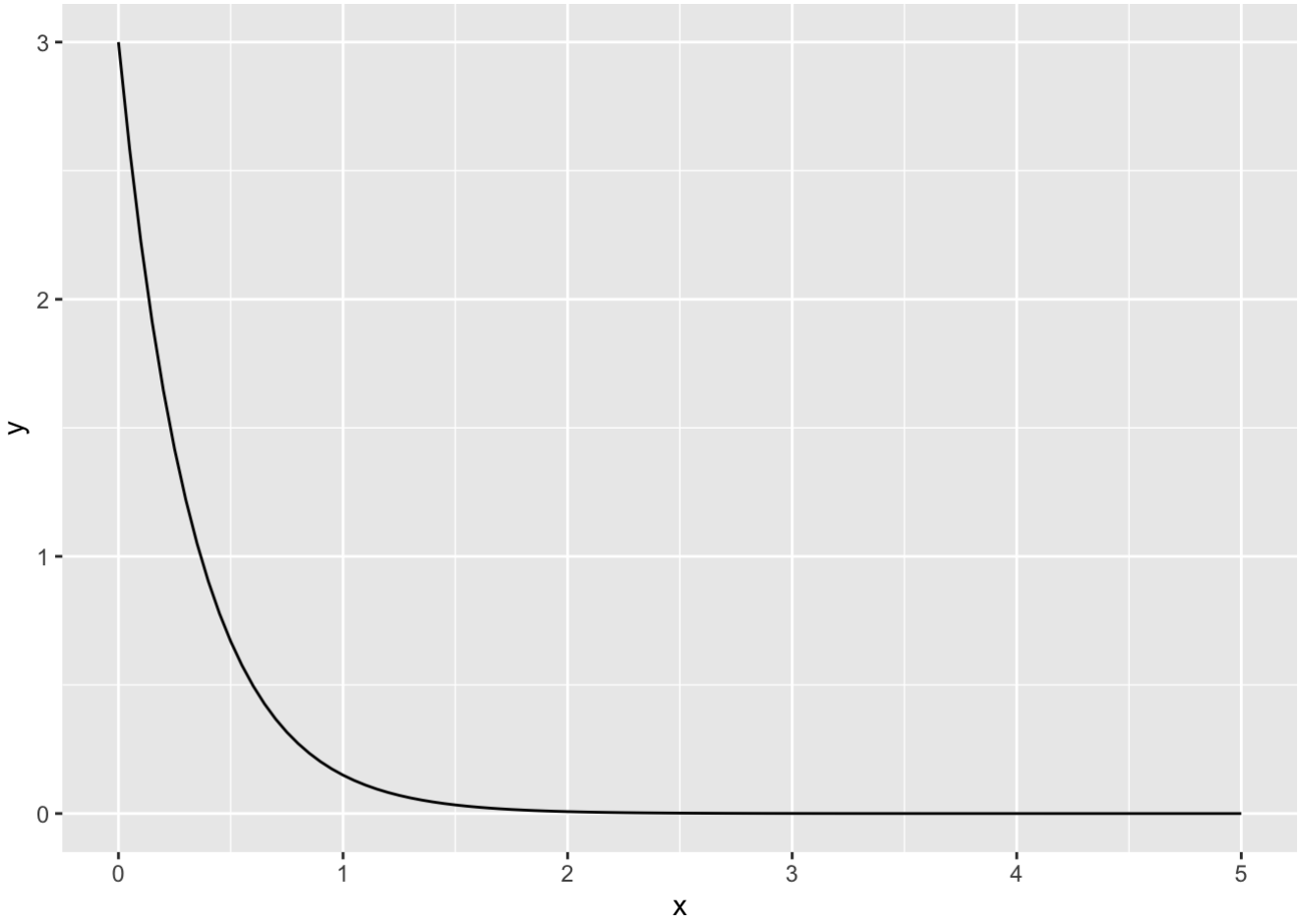
Question 2:

```
median_price <- median(craigslist$price)

below_median <- subset(craigslist, price < median_price)

num_under_2000 <- sum(below_median$price < 2000)

total_under_median = length(below_median$price)

prob= num_under_2000 / total_under_median

print(paste("Estimated Probability:", round(prob, 4)))
```

```
[1] "Estimated Probability: 0.2292"
```

Simulating with a Gamma Distribution: Question 2

```
library(ggplot2)

ggplot(data.frame(x = c(0, 5)), aes(x = x)) +
  geom_function(fun = function(x) dgamma(x, shape = 1, rate = 3))
```



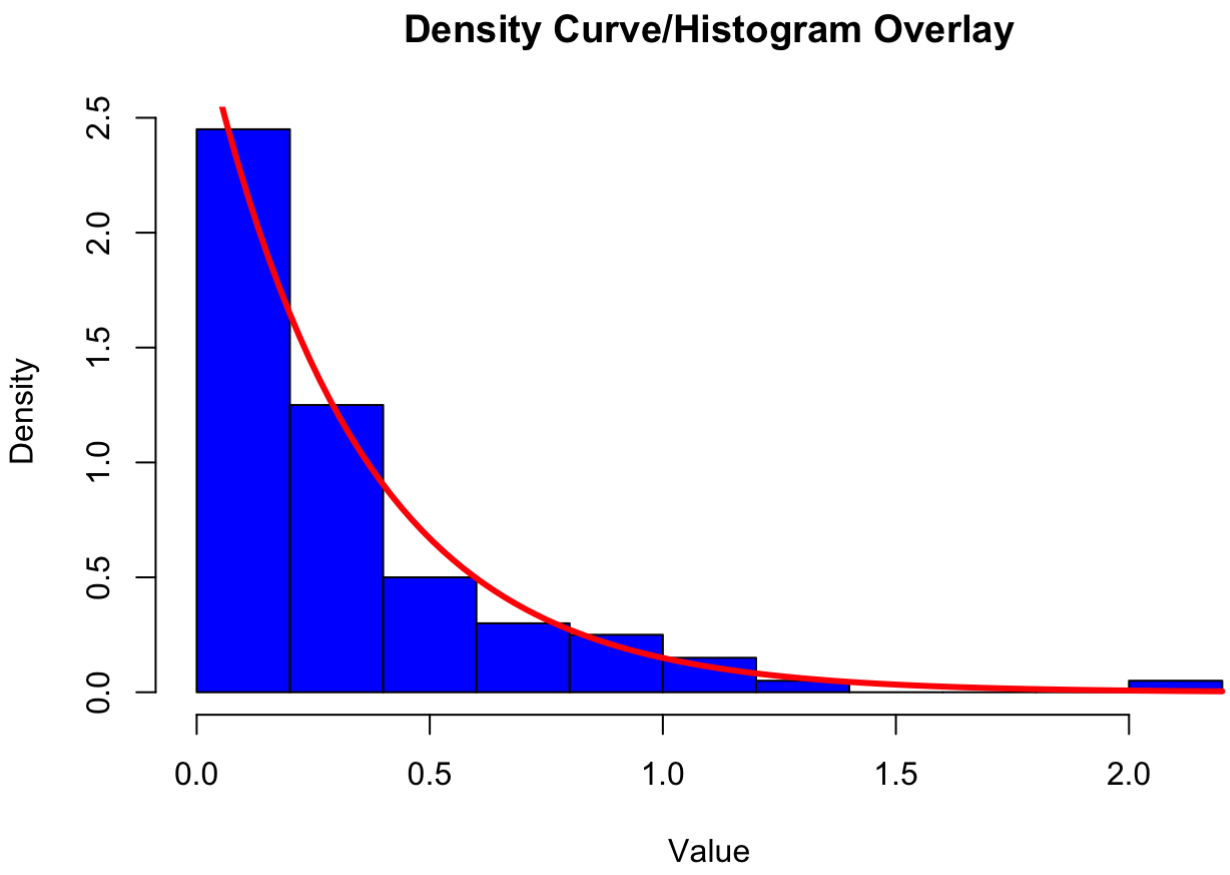
Question 4:

```
Pless_than_0_1 <- pgamma(0.1, shape = 1, rate = 3)

Pgreater_1_5 <- 1 - pgamma(1.5, shape = 1, rate = 3)

total_prob <- Pless_than_0_1 + Pgreater_1_5

rounded_prob <- round(total_prob, 4)

print(paste("Probability of observation being less than 0.1 or greater than 1.5:", rounde
```

```
[1] "Probability of observation being less than 0.1 or greater than 1.5: 0.2703"
```

Question 5:

```
simulation <- rgamma(100, shape = 1, rate = 3)

hist(simulation, freq = FALSE, main = "Density Curve/Histogram Overlay",
     xlab = "Value", ylab = "Density", col = "blue")

curve(dgamma(x, shape = 1, rate = 3), add = TRUE, col = "red", lwd = 3)
```



Question 6:

```
count <- sum(simulation < 0.1 | simulation > 1.5)

hyp_prob <- count / length(simulation)


print(paste("Estimated probability:", round(hyp_prob, 4)))
```

```
[1] "Estimated probability: 0.32"
```

Question 7

```
print("looking at our given parameters and our sample of 100 values:

Probability of an observation < 0.1:
```

```
With a sample of 100, this would be a reliable estimate as it would fall towards the apex

Probability of an observation between 0.5 and 1.0:
Probability would be better estimated as the probability of an observation between 0.5 an

Probability of an observation > 1:
As this strays away from the higher density areas of our distribution, we would need more
```

```
[1] "looking at our given parameters and our sample of 100 values:\n\nProbability of an observation < 0.1: \nWith a sample of 100, this would be a reliable estimate as it would fall towards the apex of our distribution \n\nProbability of an observation between 0.5 and 1.0:\nProbability would be better estimated as the probability of an observation between 0.5 and 1.0 would fall where the density is of highest in our sample \n\nProbability of an observation > 1: \nAs this strays away from the higher density areas of our distribution, we would need more data as we approach the tails of the distribution"
```

Distributions of Sample Data: Question 1

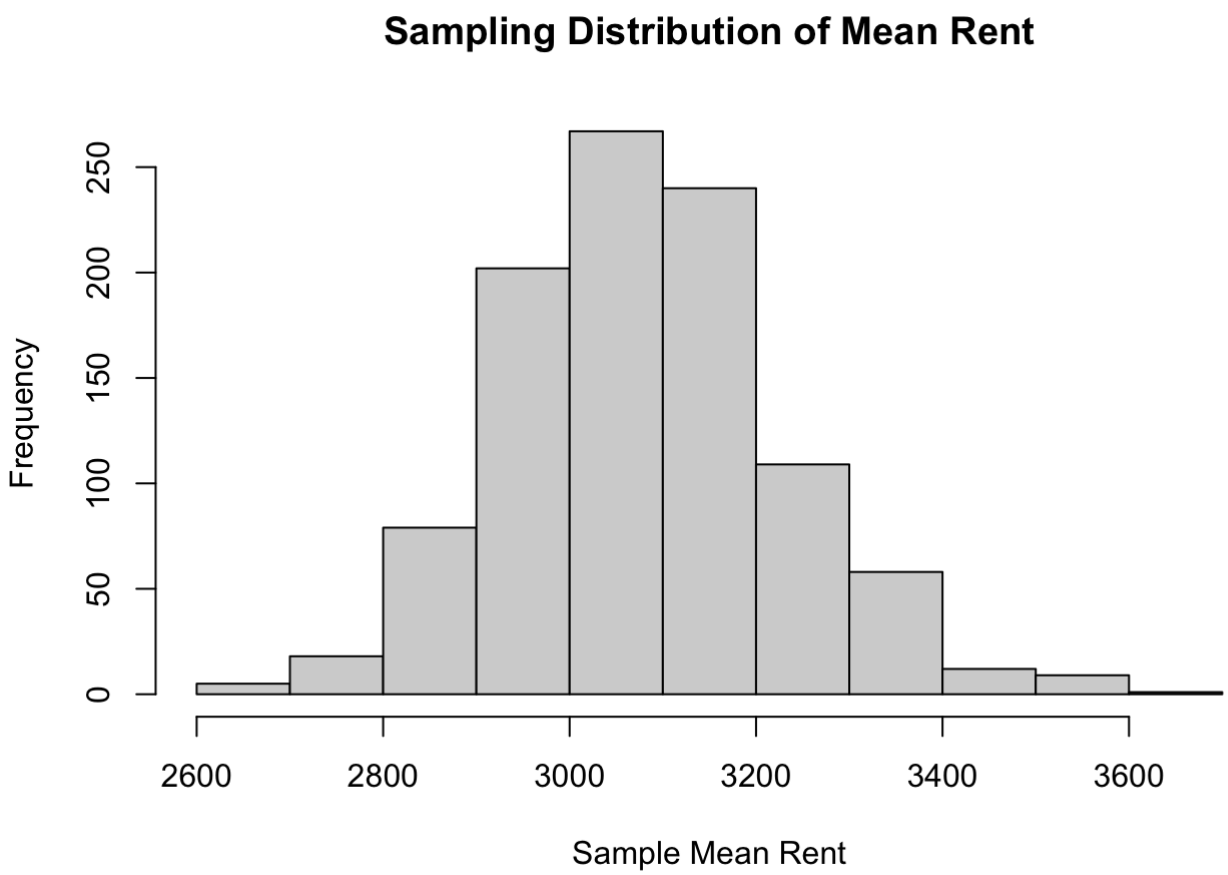
```
craigslist_all <- read.csv("~/jp-garcia-131a/craigslist_all.csv")

library(dplyr)

four_bed_craigslist_all <- craigslist_all %>%
  filter(brs <= 4)

craigslist <- four_bed_craigslist_all %>%
  sample_n(5876)

head(craigslist)
```

```
                time price size brs
1 2016-09-29 09:45:00  2262  740   1
2 2016-09-26 12:37:00  4295 1910   4
3 2016-10-08 17:40:00  2229  783   1
4 2016-09-27 08:22:00  5500 1600   3
5 2016-10-09 13:11:00  3500 1328   3
6 2016-09-26 17:49:00  2250  640   1
                                                    title
1         Enjoy The Best Deals In San Jose 8 Weeks Free  99 Deposit
2    Cambrian Beauty Sunny and Bright 4BR 3BA KB Single Family Home
3      READY TO MOVE TODAY OneBedroom Apartment Home Turn Key Ready
4 Modern Luxury Townhome in Heart of Mill Valley Available for Rent
5                                                Condo for rent
6     Upgraded Washer  Dryer Pet Friendly And First Month Rent Free
                      link                        location
1 /sby/apa/5804920120.html                 san jose south
2 /sby/apa/5764055699.html          willow glen / cambrian
3 /eby/apa/5819684163.html dublin / pleasanton / livermore
4 /nby/apa/5789034890.html                    mill valley
```

```
5 /eby/apa/5820624683.html           danville / san ramon
6 /sby/apa/5800823085.html                 san jose north
```

Question 2:

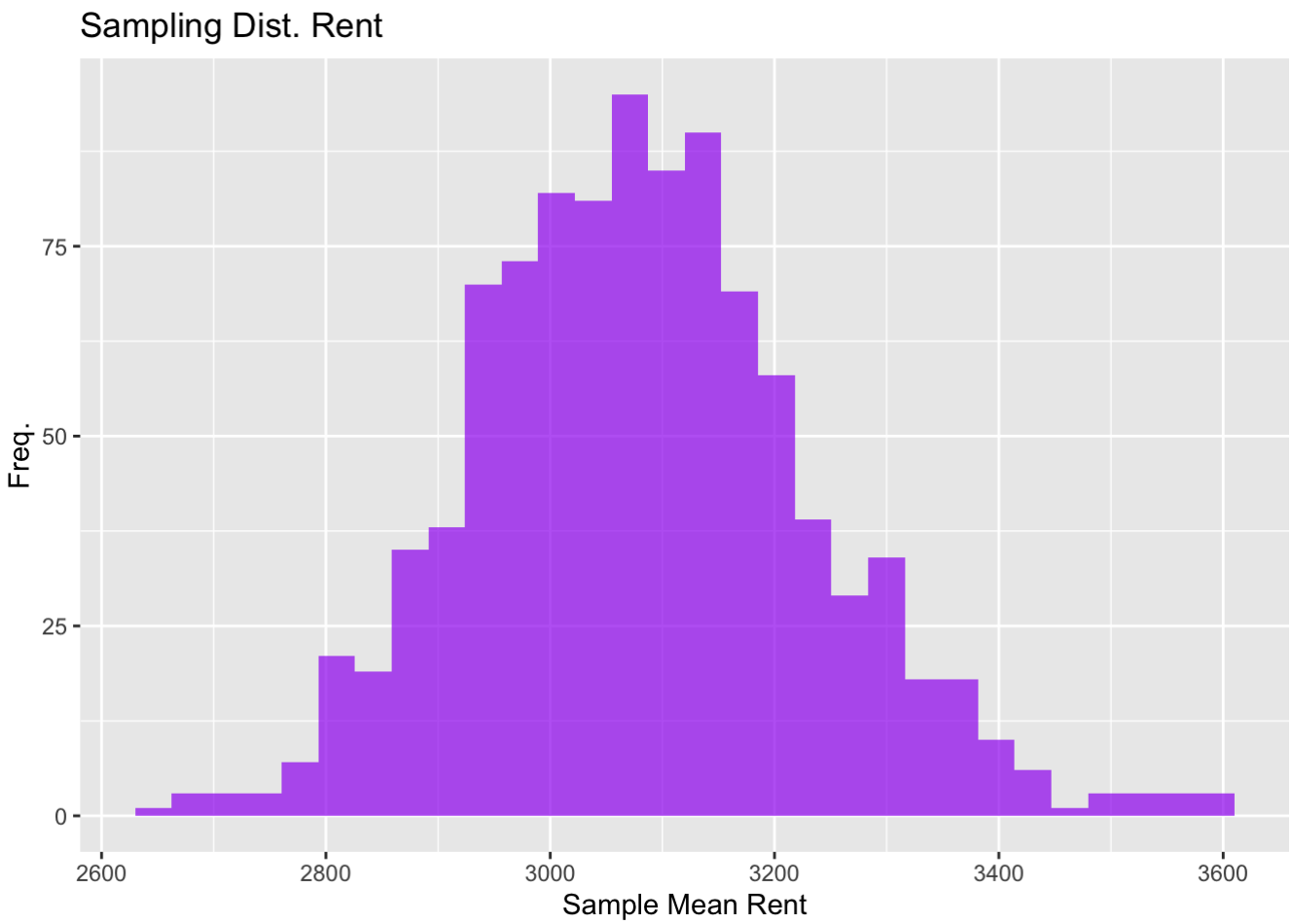
```r
# Sample size
n <- 100

# Number of samples
samples <- 1000

sample_means <- replicate(samples, {
  sample_data <- sample_n(subset(craigslist_all, brs <= 4), n)
  mean(sample_data$price)
})

# Visualize the sampling distribution
hist(sample_means, main = "Sampling Distribution of Mean Rent", xlab = "Sample Mean Rent"
```



Question 3:

```r
library(ggplot2)
sample_means_df <- data.frame(sample_means)
```

```
ggplot(sample_means_df, aes(x = sample_means)) +
  geom_histogram(fill = 'purple', alpha = 0.7, bins = 30) +
  labs(title = "Sampling Dist. Rent", x = "Sample Mean Rent", y = "Freq.")
```



Question 4:

```
print("Extremely unsure how to go about this problem")
```

```
[1] "Extremely unsure how to go about this problem"
```

Question 5:

```
x_vector <- c(0, 1, -2)  # Possible values
probs_vector <- c(1/3, 1/3, 1/3)  # Associated probabilities

# Number of 'X' to sum for each observation (sample size)
m <- 10000

# Number of observations (replications)
B <- 5000

n_sums <- replicate(
  B,
```

```
  mean(
    sample(x = x_vector, size = m, replace = TRUE, prob = probs_vector)
  )
)
```

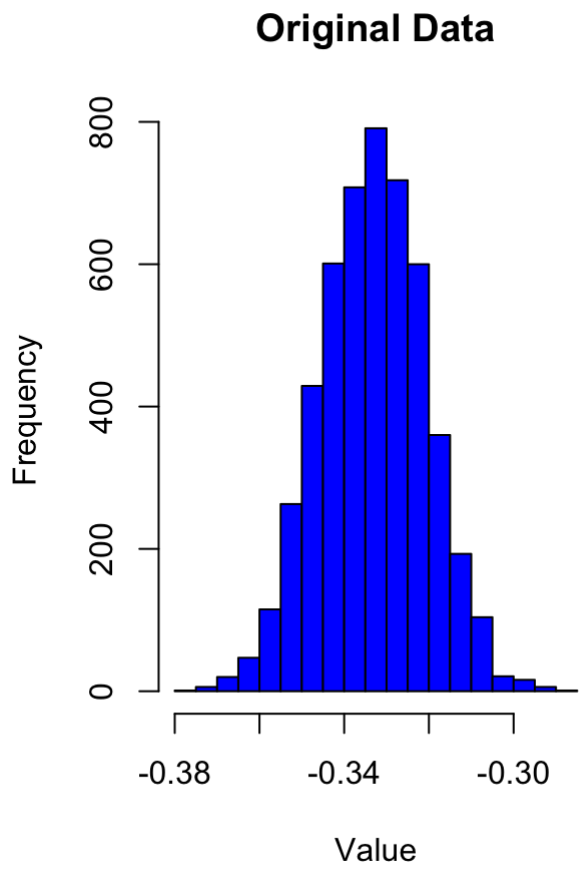
Question 6:

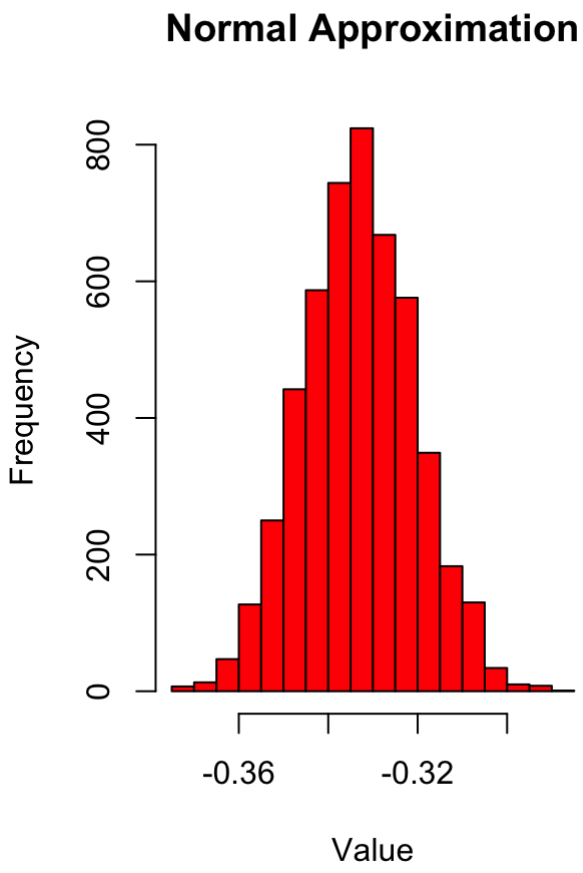
```
mu <- -1/3
variance <- 14/9 / 10000
sd <- sqrt(variance)
n_sum_normal <- rnorm(5000, mean = mu, sd = sd)

summary(n_sum_normal)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.3737 -0.3418 -0.3334 -0.3332 -0.3248 -0.2887 
```

Question 7:

```
# Create a new window for multiple plots
par(mfrow=c(1,2))

# Histogram for original data
hist(n_sums, main="Original Data", xlab="Value", ylab="Frequency", col="blue")

# Histogram for approximate data
hist(n_sum_normal, main="Normal Approximation", xlab="Value", ylab="Frequency", col="red"
```





Density Estimation: Question 1

```
library(dplyr)

heart_disease <- read.csv("~/jp-garcia-131a/heartDisease.csv")

library(dplyr)

# Assuming heartDisease is your data frame
heart_disease <- heart_disease %>%
  mutate(
    num = factor(num),
    cp = factor(cp)
  )
str(heart_disease)
```

```
'data.frame':   297 obs. of  14 variables:
 $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
 $ sex     : num  1 1 1 1 0 1 0 0 1 1 ...
 $ cp      : Factor w/ 4 levels "1","2","3","4": 1 4 4 3 2 2 4 4 4 4 ...
 $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
 $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
 $ fbs     : num  1 0 0 0 0 0 0 0 0 1 ...
 $ restecg : num  2 2 2 0 2 0 2 0 2 2 ...
```

```
 $ thalach : num  150 108 129 187 172 178 160 163 147 155 ...
 $ exang   : num  0 1 1 0 0 0 0 1 0 1 ...
 $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
 $ slope   : num  3 2 2 3 1 1 3 1 2 3 ...
 $ ca      : num  0 3 2 0 0 0 2 0 1 0 ...
 $ thal    : num  6 3 7 3 3 3 3 3 7 7 ...
 $ num     : Factor w/ 5 levels "0","1","2","3",..: 1 3 2 1 1 1 4 1 3 2 ...
```

Question 2:

```
heart_disease %>%
    select(num,cp) %>%
    table()
```

```
   cp
num  1  2  3  4
  0 16 40 65 39
  1  5  6  9 34
  2  1  1  4 29
  3  0  2  4 29
  4  1  0  1 11
```

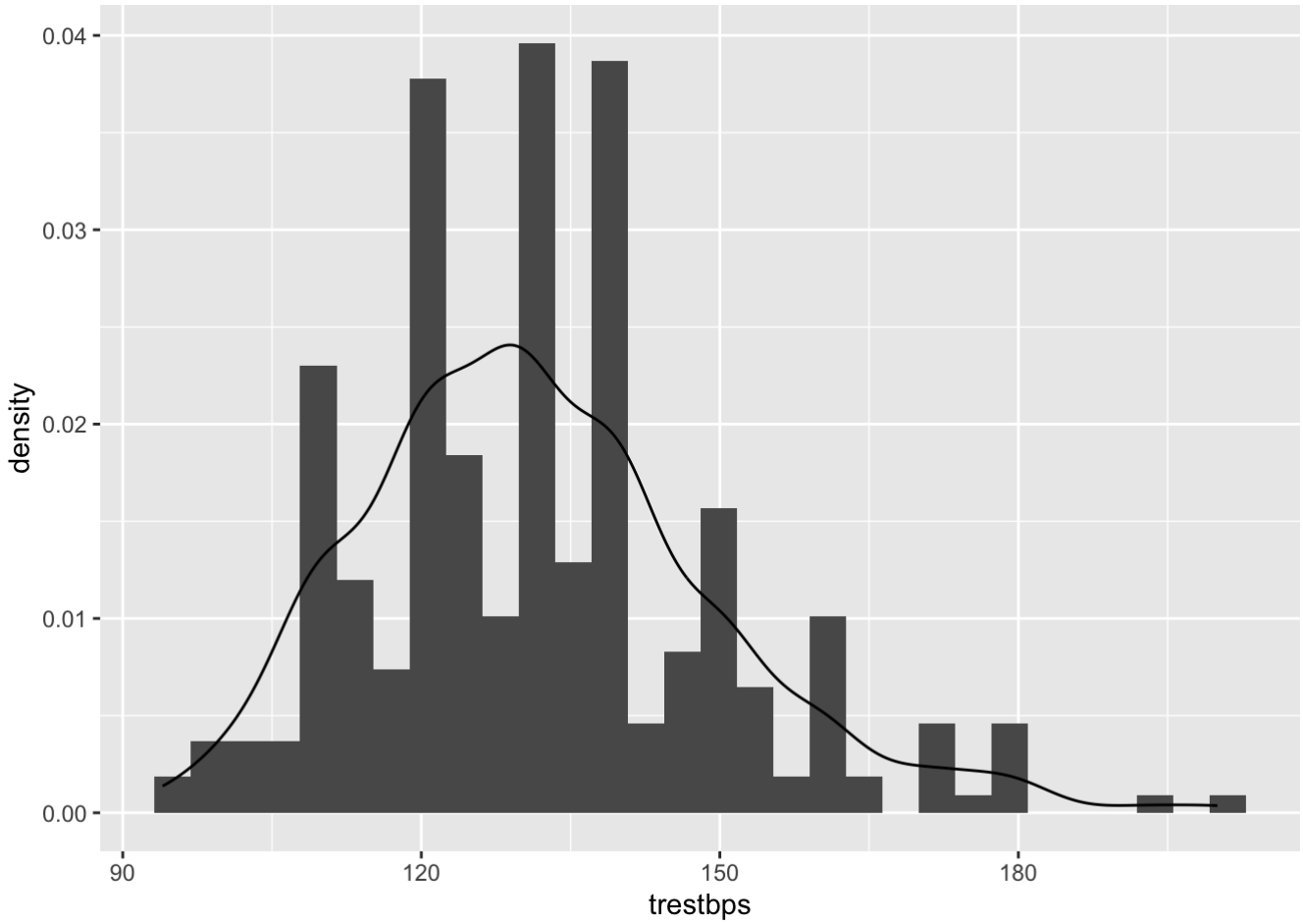
```
print("This code will only select the num and cp columns from the heart_disease data fram
```

```
[1] "This code will only select the num and cp columns from the heart_disease data frame
and the table function will give us a table summary of how many occurences there are per
unique combination of num and cp"
```

Question 3:

```
library(ggplot2)

ggplot(heart_disease, aes(x=trestbps)) +
  geom_histogram(aes(y = ..density..), bins = 30) +
  geom_density()
```

```
Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
ℹ Please use `after_stat(density)` instead.
```

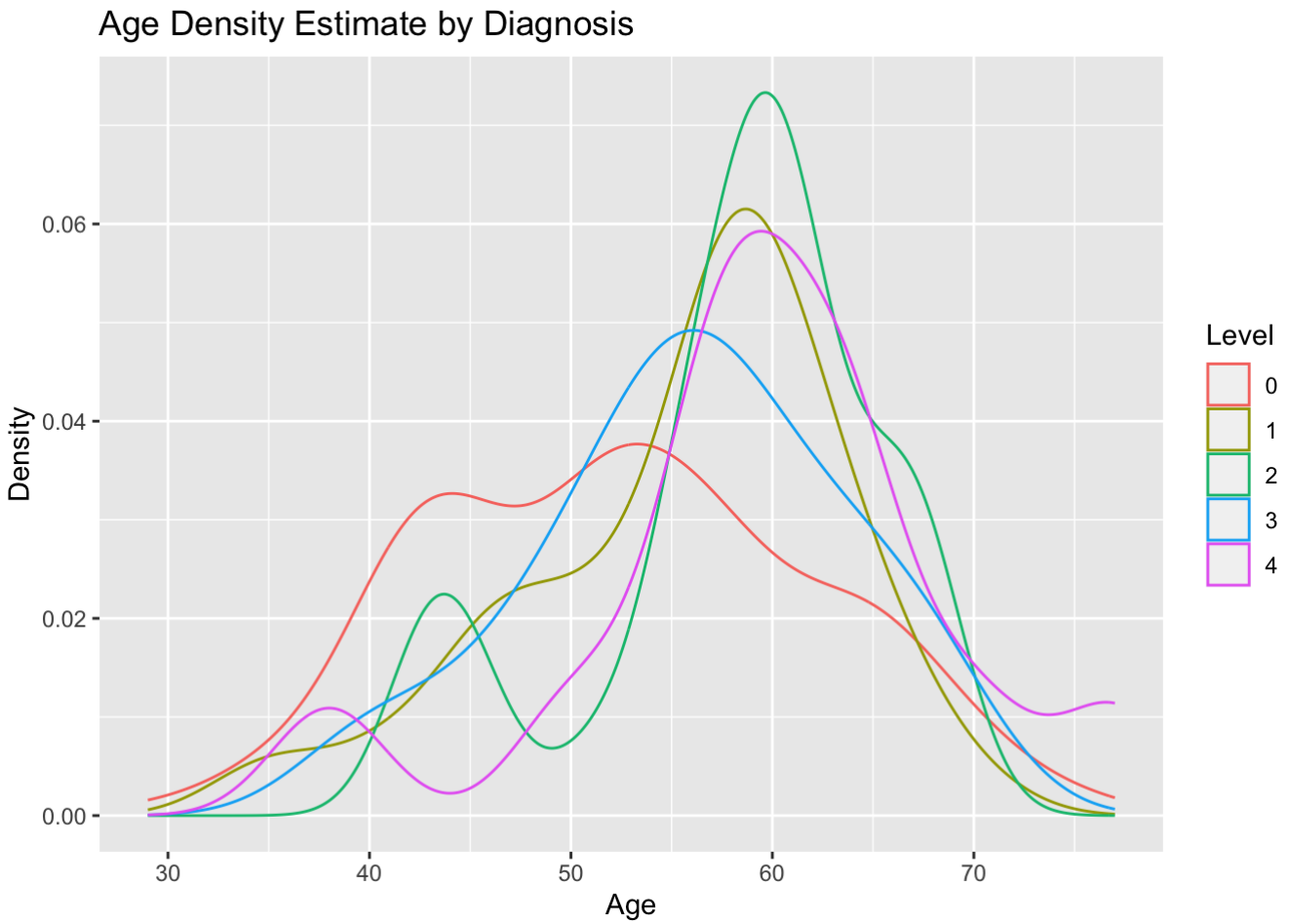



```
print("Looking at the shape of the distribution we are able to see that there is a right
```

```
[1] "Looking at the shape of the distribution we are able to see that there is a right
skewed but largely centered around values 120-125. The blood pressure seems to be mostly
normal while there is a proportion that seems to hover in the high pressure"
```

Question 4:

```
library(ggplot2)

ggplot(heart_disease, aes(x = age, color = factor(num))) + #Different colored line for ea
  geom_density() +
  labs(title = "Age Density Estimate by Diagnosis",
       x = "Age",
       y = "Density",
       color = "Level")
```


Age Density Estimate by Diagnosis
Density
0.06
0.04
0.02
0.00
30
40
50
60
70
Age
Level
0
1
2
3
4