STAT 5650 Statistical Learning and Data Mining 1   Spring 2020

Names: Greg Hoffmann, Kristen Sohm, Michael Huber, Varsha Reddy Mandadi

Submission Date: 04/22/2020

Final Project: Applying Prediction Methods to Hotel Data

200 Points — Due Friday 10/18/2019 (via Canvas by 11:59pm)

(i) **Prediction Method 1:** After working with the data and reviewing the notes from Dr. Cutler on our project proposal we had to edit what methods of prediction we were going to apply against the data. Below is a list of methods that we were able to get to run against our hotel data set.

    (a) (1 Point) Load all required R packages to answer this question. Show your R code.
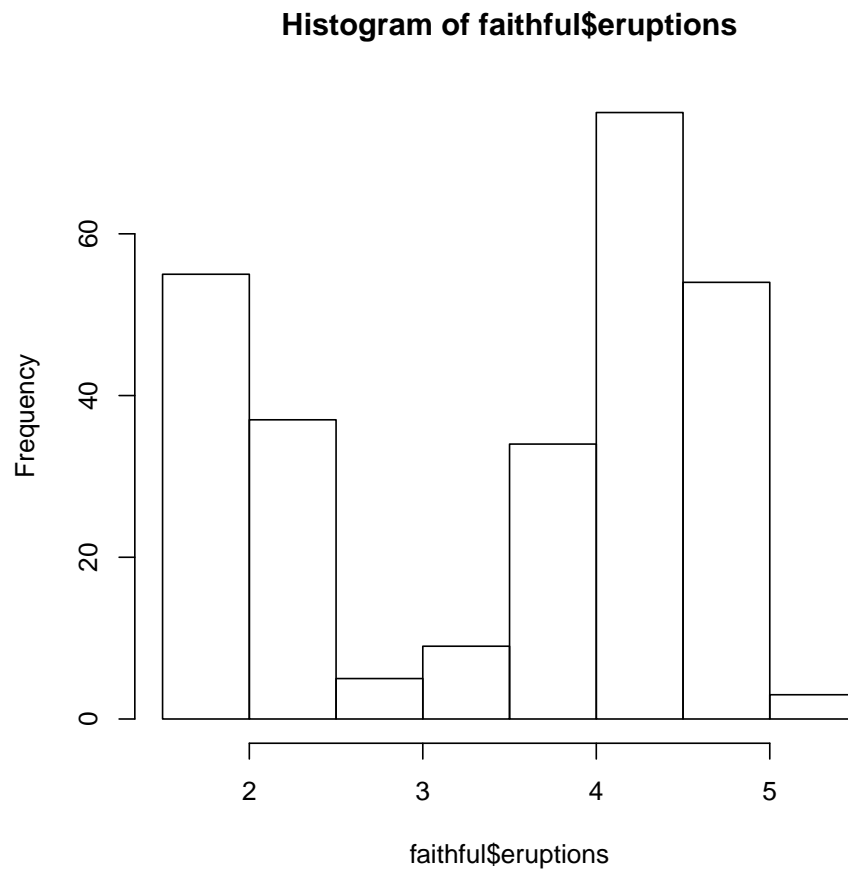
    (b) (1 Point) <first requirement with code>
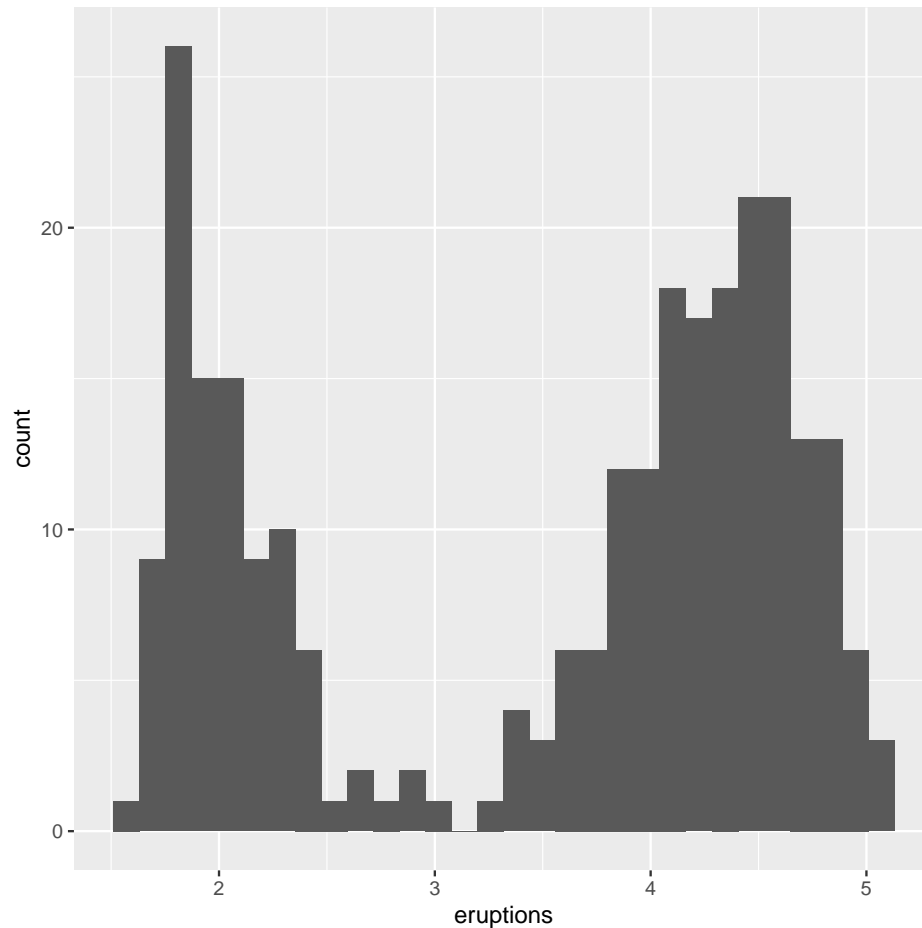
    Answer:

- <Answer Part 1>
- <Answer Part 2>

   (c) (1 Point) Repeat (b) from above, now using the *hist* function from baseR.

```
> hist(faithful$eruptions)
```

2

Histogram of faithful$eruptions

(d) (3 Points) Repeat (c) from above, now using the *hist* function from baseR.

```
> hist(faithful$eruptions,
+       breaks = seq(1, 6, .5),
+       xlab = "Eruption Time in Minutes",
+       ylab = "Count of Eruptions",
+       main = "Old Faithful Eruption Data Histogram",
+       col = "light blue",
+       las = 1,
+       ylim = c(0, 100),
+       xlim = c(1, 6))
```
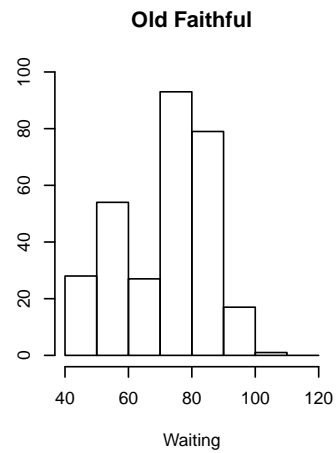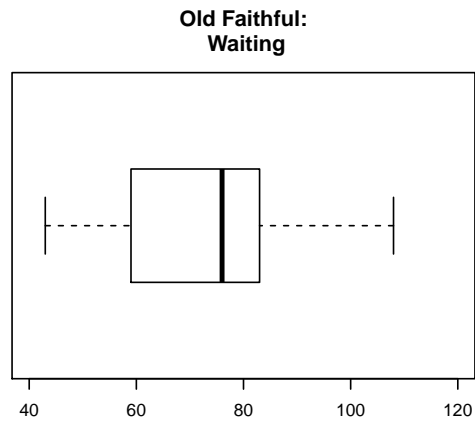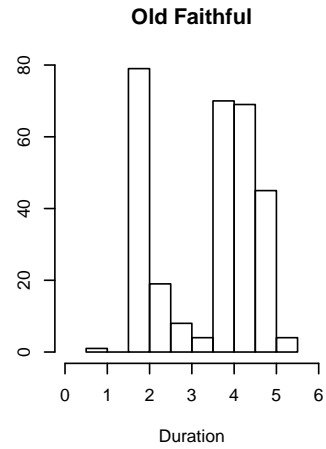
Old Faithful Eruption Data Histogram

Count of Eruptions

Eruption Time in Minutes
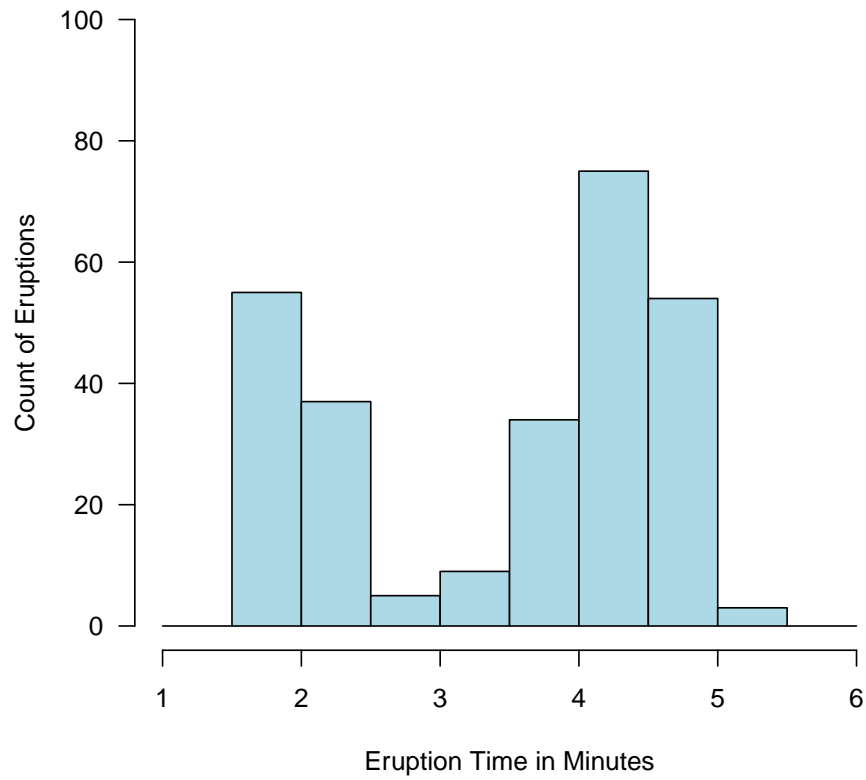
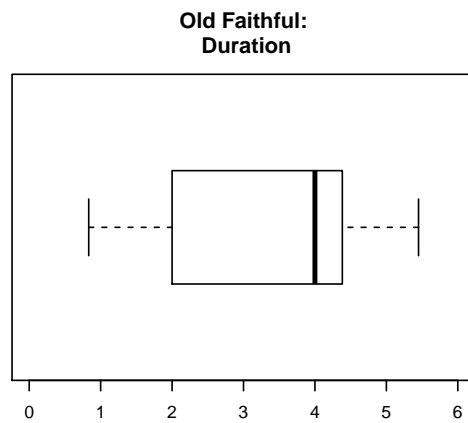Answer:

- \<Answer Part 1\>
- \<Answer Part 2\>

(ii) **Question 2:** <Description of Question 2>

(a) (6 Points) Recreate the graphs (and layout) below using baseR. Use a ruler to check that the width and height proportions in your graphs match the ones I have used. I worked with integer multiples! Include your R code and the resulting graphs. Hint: You can create a new line via \n without any extra spaces before/after \n.

```
> grid <- matrix(c(1, 1, 1, 2, 2, 3, 3, 3, 4, 4),
+                nrow = 2, ncol = 5, byrow = TRUE)
> layout(grid)
> par(mar = c(4, 3, 4, 2))
> boxplot(geyser$duration,
+         horizontal = TRUE,
+         main = "Old Faithful:\n Duration",
+         ylim = c(0, 6))
> hist(geyser$duration,
+      main = "Old Faithful",
+      xlab = "Duration",
+      ylab = "Count",
+      xlim = c(0, 6))
> boxplot(geyser$waiting,
+         horizontal = TRUE,
+         main = "Old Faithful:\n Waiting",
+         ylim = c(40, 120))
> hist(geyser$waiting,
+      main = "Old Faithful",
+      xlab = "Waiting",
+      ylab = "Count",
+      xlim = c(40, 120),
+      ylim = c(0, 100),
+      breaks = seq(40, 120, by = 10))
```

**Old Faithful:
Duration**

Old Faithful

Duration

**Old Faithful:
Waiting**

Old Faithful

Waiting

Refernces:

- https://www.statmethods.net/advgraphs/layout.html
- https://stackoverflow.com/questions/31319942/change-the-size-of-a-plot-when-plotting-multiple-plots-in-r
- https://www.youtube.com/watch?v=Z3V4Pbxeahg

(b) (2 Points) Recreate the graph below using ggplot2. Include your R code and
the resulting graph.

```
> ggplot(geyser, aes(x=duration, y=waiting)) +
+    geom_point() +
+    xlab("Duration") +
+    ylab("Waiting") +
+    xlim(0, 6) +
+    ylim(40, 120) +
+    ggtitle("Old Faithful Data") +
+    theme(plot.title = element_text(hjust = 0.5))
```

Old Faithful Data
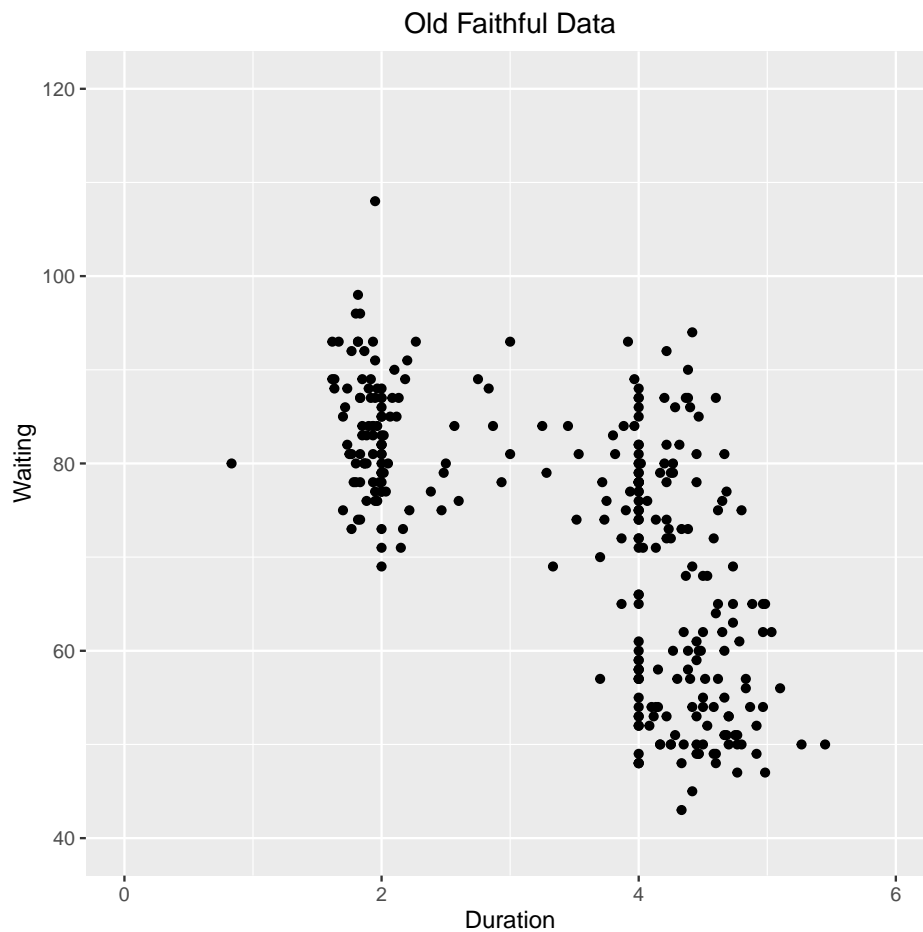
(c) (2 Points) Doesn't the scatterplot in (b) above look rather different than
the scatterplot in Question 1 (j)? Note that the help page for *geyser* states
`waiting   numeric   Waiting time for this eruption` and
`The waiting time was incorrectly described as the time to the next`
`eruption in the original files, and corrected for MASS version 7.3-30.`
Use this information to create a basic scatterplot for the *geyser* data that
matches the overall appearance in Question 1 (j). Include your R code and
the resulting graph. No need to refine this scatterplot.

```
> duration <- geyser$duration[1:298]
> waiting <- geyser$waiting[2:299]
> df <- data.frame(duration, waiting)
> ggplot(df, aes(x = duration, y = waiting )) +
+           geom_point()
```
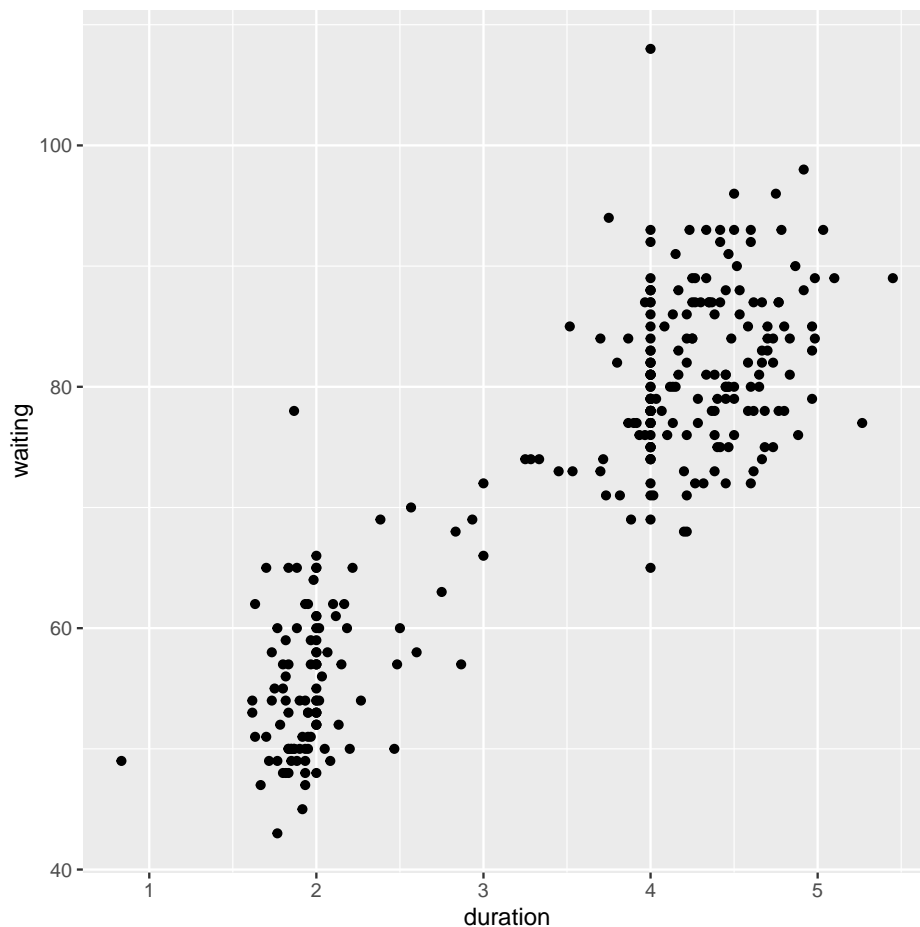

Answer: The geyser data appears to have a negative correlation with three
clusters where the faithful data has a positive correlation with 2 clusters.

General Instructions

(i) <Instruction 1>

(ii) <Instruction 2>