STAT 33A Workbook 6

CJ HINES (3034590053)

Oct 8, 2020

This workbook is due **Oct 8, 2020** by 11:59pm PT.

The workbook is organized into sections that correspond to the lecture videos for the week. Watch a video, then do the corresponding exercises *before* moving on to the next video.

Workbooks are graded for completeness, so as long as you make a clear effort to solve each problem, you'll get full credit. That said, make sure you understand the concepts here, because they're likely to reappear in homeworks, quizzes, and later lectures.

As you work, write your answers in this notebook. Answer questions with complete sentences, and put code in code chunks. You can make as many new code chunks as you like.

In the notebook, you can run the line of code where the cursor is by pressing `Ctrl` + `Enter` on Windows or `Cmd` + `Enter` on Mac OS X. You can run an entire code chunk by clicking on the green arrow in the upper right corner of the code chunk.

Please do not delete the exercises already in this notebook, because it may interfere with our grading tools.

You need to submit your work in two places:

- Submit this Rmd file with your edits on bCourses.
- Knit and submit the generated PDF file on Gradescope.

If you have any last-minute trouble knitting, **DON'T PANIC**. Submit your Rmd file on time and follow up in office hours or on Piazza to sort out the PDF.

Exploratory Data Analysis

Watch the "Exploratory Data Analysis" lecture video.

No exercises for this section. Almost done!

Distribution Plots

Watch the "Distribution Plots" lecture video.

Exercise 1

1. Use the Dogs data set and the ggplot2 package to create a density plot. The plot should show the distribution of weights grouped by the grooming needs of the breed.

2. Use the ggridges package to create a ridges plot that shows the same information as the plot from part 1.

3. In 2-5 sentences, comment on if and how weights relate to grooming needs.

YOUR ANSWER GOES HERE:

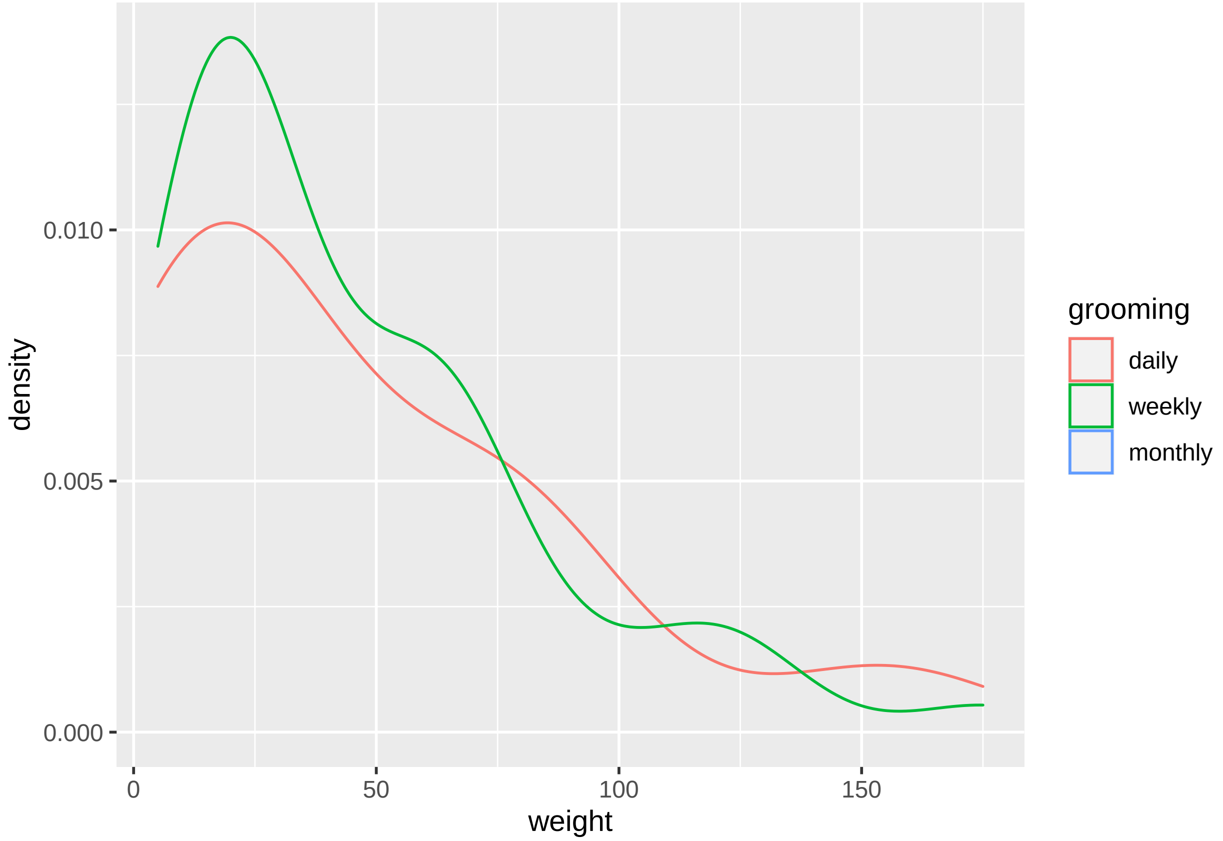
1.

```
dogs = readRDS("dogs.rds")
library(ggplot2)
dogs2 = dogs[!is.na(dogs$grooming),]
ggplot(dogs2, aes(x = weight, color = grooming)) + geom_density()
```

```
## Warning: Removed 47 rows containing non-finite values (stat_density).
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```

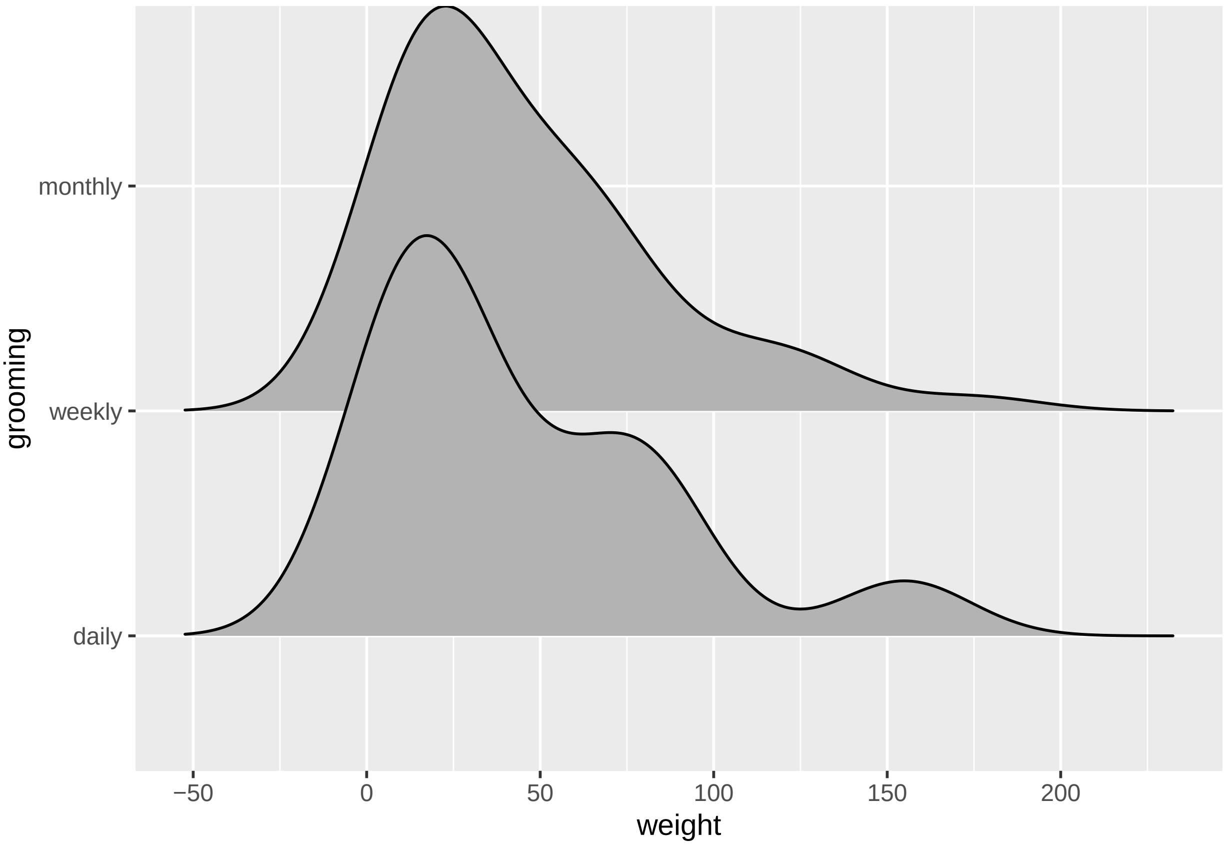


2.

```r
library(ggridges)
ggplot(dogs2, aes(x = weight, y = grooming)) + geom_density_ridges()
```

```
## Picking joint bandwidth of 19.1
```

```
## Warning: Removed 47 rows containing non-finite values (stat_density_ridges).
```



3.

There's a spike showing that dogs below 75 pounds are more likely to be groomed weekly rather than daily. Grooming needs do not really vary with weight, especially in dogs with larger weights. Dogs of all sizes are groomed either daily or weekly.

Faceted Plots

Watch the "Faceted Plots" lecture video.

Exercise 2

1. Use the Datasaurus Dozen data set to create a faceted scatter plot. Shows each dataset from the Datasaurus Dozen in a separate facet. Use `geom_smooth()` with `method = "lm"` to add a linear regression line to each facet.

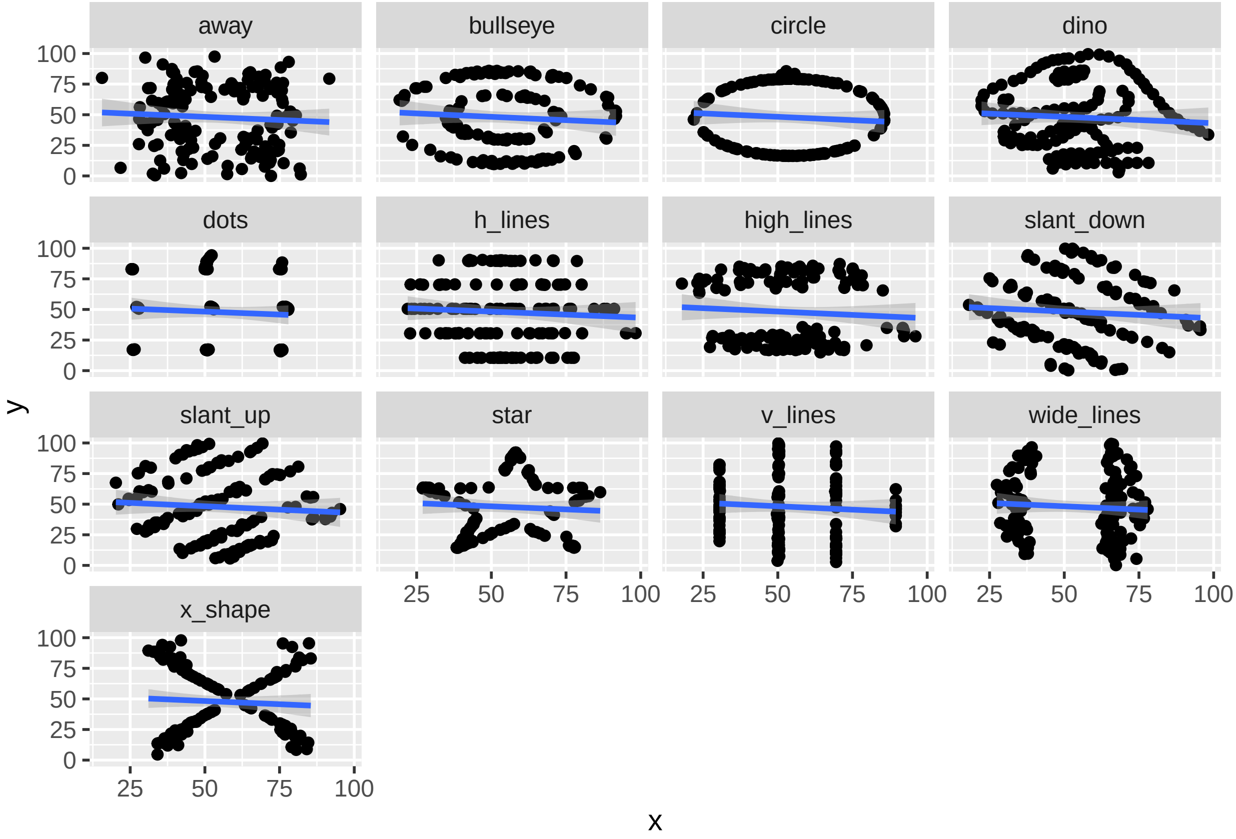
2. Is there any pattern to the regression lines across the different data sets?

YOUR ANSWER GOES HERE:

1.

```
dino = read.delim("../hws/DatasaurusDozen.tsv")
ggplot(dino, aes(x = x, y = y)) + geom_point() + facet_wrap(vars(dataset))  + geom_smooth(method = lm)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



2. The regression lines are horizontal at 50 on the y-axis. They all seem to be the same since the plots aren’t linear.

EDA Strategy

Watch the “EDA Strategy” lecture video.

No exercises for this section. Almost done!

EDA Examples

Watch the “EDA Examples” lecture video.

Exercise 3

1. Come up with 2 questions it might be possible to answer using the Craigslist Apartments data. Think about what the columns in the data set are actually capable of answering. Avoid simple questions that are likely to yield a yes or no answer with minimal investigation–these usually aren't interesting. *Hint: Briefly inspect the data set to see what's there before coming up with questions.*
2. Make a plot or compute statistics to help answer your first question. Discuss your result in 1-3 sentences. It's okay if you don't haven't completely "solved" the question as long as you're able to meaningfully address what it's asking.
3. Repeat part 2 for you second question.

YOUR ANSWER GOES HERE:

1.

   Q1: How is sqft distributed in Los Angeles, Long Beach, and Lancaster?

   Q2: How many bathrooms are typically in Berkeley apartment listings?

2. How is square footage distributed in Los Angeles, Long Beach, and Lancaster? More of the listings are in Los Angeles rather than Long Beach and Lancaster, so the plot is skewed with the tail to the right for outlier listings over 2500 square feet in Los Angeles. In Lancaster, the majority of listings are less than 1250 square feet. Below 2500 square feet, Long Beach and Los Angeles listings follow a similar distribution that peak around 1000 square feet and gradually fall in frequency.

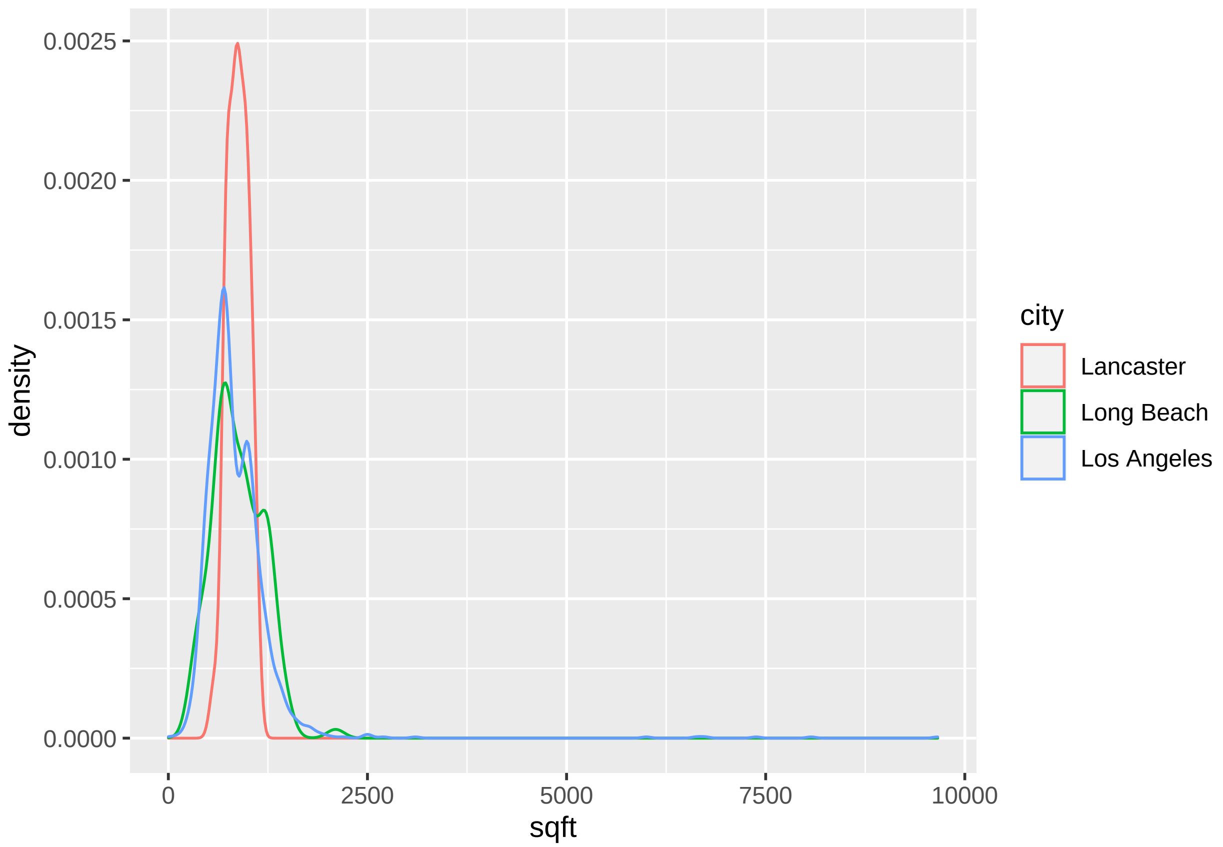
```r
apts = readRDS("Apartments.rds")
#str(apts)

Lcities = c("Los Angeles", "Long Beach", "Lancaster")
Litties = apts[apts$city %in% Lcities, ]
Litties$city = droplevels(Litties$city)
table(Litties$city, useNA = "always")
```

```
##
##   Lancaster  Long Beach Los Angeles        <NA>
##          40         120        1890           0
```

```r
ggplot(Litties, aes(sqft, color = city)) + geom_density()
```

```
## Warning: Removed 435 rows containing non-finite values (stat_density).
```



3. How many bathrooms are typically in Berkeley apartment listings? Most apartments in Berkeley are one bathroom apartments and the second most common are two bathroom apartments. There are apartments with 3 or 4 bathrooms which I found surprising.

```
apts = readRDS("Apartments.rds")
berkeley = apts[apts$city %in% "Berkeley",]
table(berkeley$bathrooms, useNA = "no")
```

```
##
##   1 1.5   2 2.5   3   4
## 416   5  43   1   2   3
```