STAT 33B Lab 3

Gunnar Mayer (3034535154)

Edit this file, knit to PDF, and:

- Submit the Rmd file on bCourses.
- Submit the PDF file on Gradescope.

If you think you'll need help with submission, please ask during the lab.

Answer all questions with complete sentences, and put code in code chunks. You can make as many new code chunks as you like. Please do not delete the exercises already in this notebook, because it may interfere with our grading tools.

As you work, you may find it helpful to be able to run your code. You can run a single line of code by pressing Ctrl + Enter. You can run an entire code chunk by clicking on the green arrow in the upper right corner of the code chunk.

Knit the document from time to time to make sure that your code runs without errors from top to bottom in a fresh R environment.

The code below controls the number of significant digits shown for the return values in your knitted document.

```r
options(digits = 3)
```

Datasaurus Dozen Data Set

Exercise 1 uses the Datasaurus Dozen Data Set from homework 2.

Exercise 1

A "faceted" plot is one that shows several plots side-by-side, to aid comparison between them. Each subplot is called a "facet".

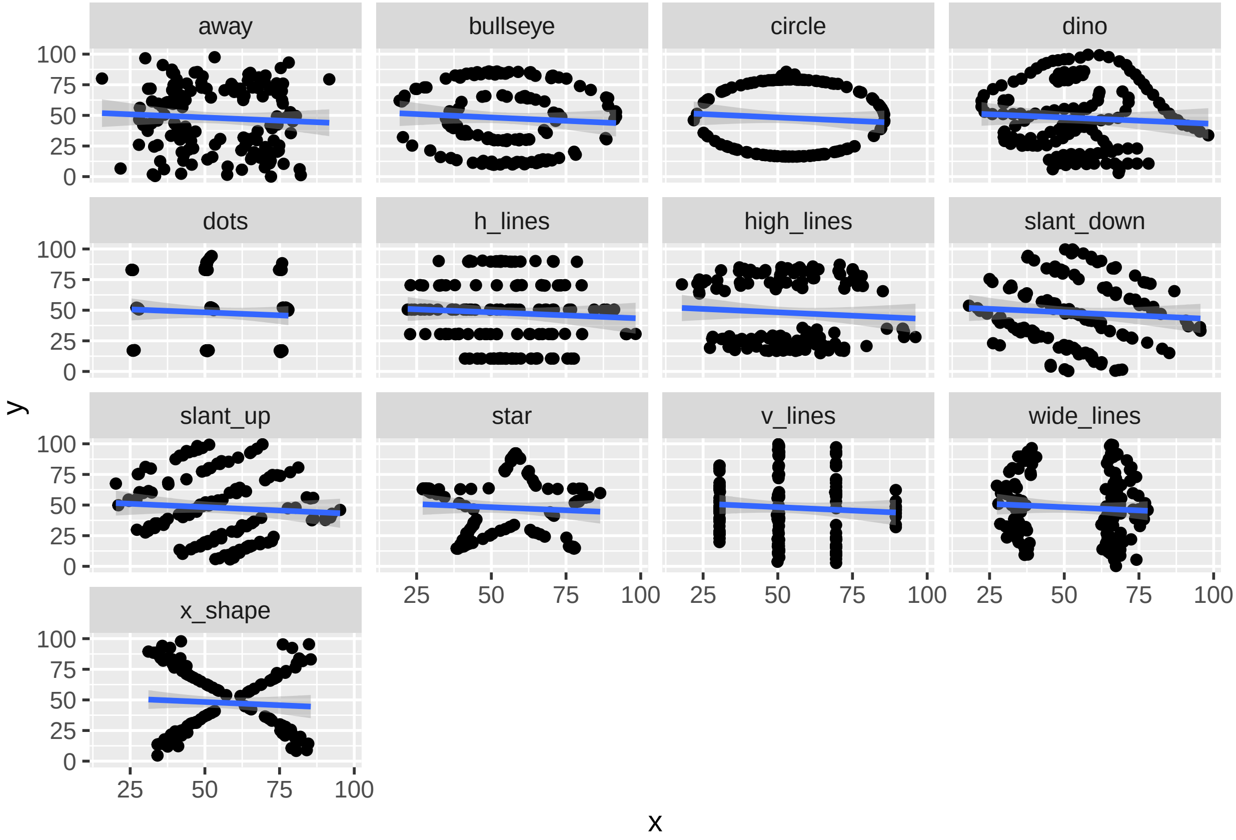
You can create a faceted plot with `ggplot2` by using the facet layer. For instance, the `facet_wrap()` function creates a line of facets based on a single categorical variable. The facet layer should be added to a plot *after* the geometry layers.

1. Read the documentation for `facet_wrap()`, then create a faceted scatter plot that 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.

   Hint: Unlike other `ggplot2` functions, the facet functions do not accept unquoted variable names. See the documentation for details.

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

```r
# Your code goes here.
library(ggplot2)
dsaur = read.delim(file = "DatasaurusDozen.tsv")
ggplot(data = dsaur, aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm") +
    facet_wrap(vars(dsaur$dataset))
```



YOUR WRITTEN ANSWER GOES HERE: Is there any pattern to the regression lines across the different data sets?

Yes there is a pattern to the regression lines across the different data sets. All of the regression lines have very similar slopes. This means that the average of all the datasets inside dsaur will be relatively similar.

Bay Area Apartments Data Set

For the remainder of this lab, you'll continue to analyze the Bay Area Apartments Data Set introduced in lab 2.

Recall that the Bay Area Apartments Data Set is a collection of 5852 advertisements for apartments for rent in the San Francisco Bay Area. The data set was collected from Craigslist on Feb 13, 2020.

Exercise 2

Histograms, density plots, and box plots are all good ways to display the distribution of a 1-dimensional numerical feature. You can use side-by-side box plots to compare a distribution across the levels of a categorical feature,

1. Use `ggplot2` to create side-by-side box plots that show the distribution of apartment prices broken down by number of bathrooms. Add descriptive labels and a title to your plot.

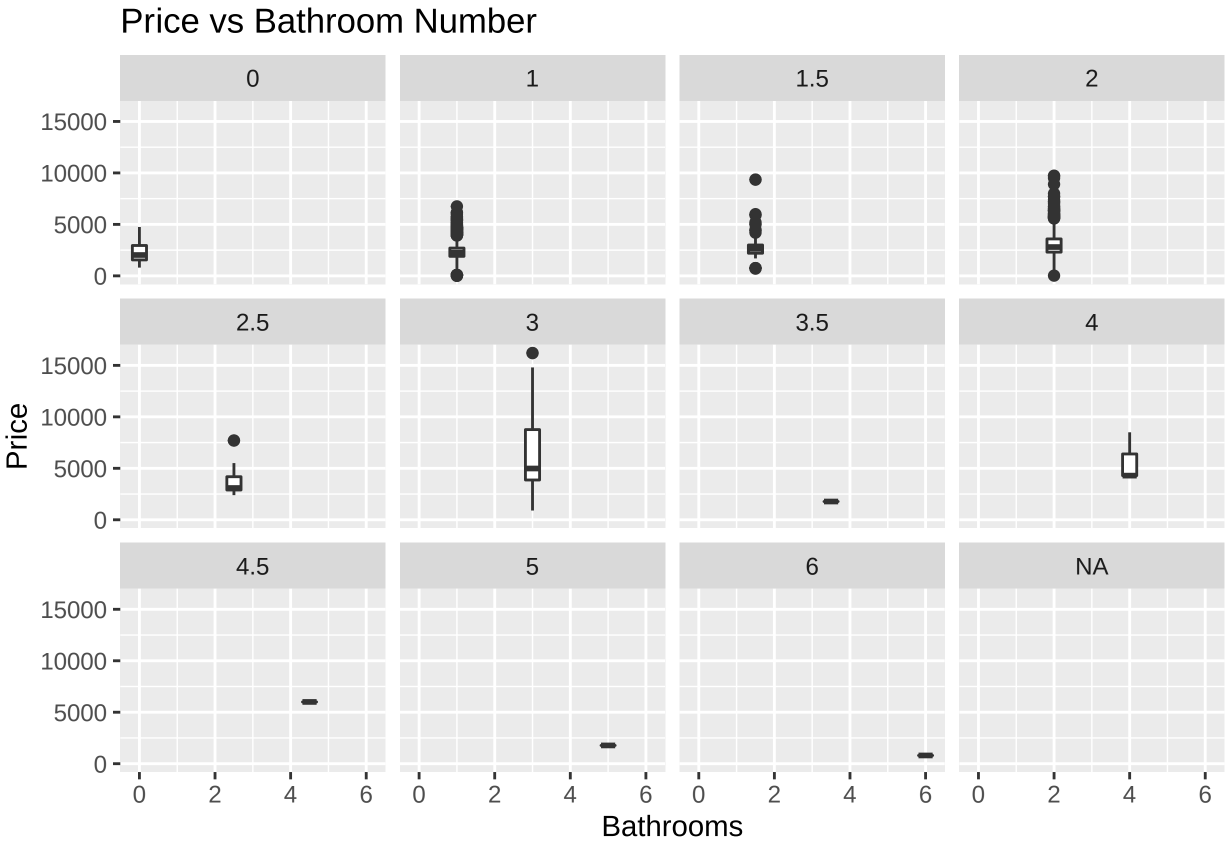
   Hint: The `ggplot2` package treats discrete data and categorical data differently. You can use the `factor()` function to convert a vector or column into categorical data.

2. Based on your plot, does the number of bathrooms affect the price? If so, how?

```
ba_data = readRDS("cl_apartments.rds")
ggplot(data = ba_data) +
    geom_boxplot(aes(x = bathrooms, y = price)) +
    facet_wrap(vars(ba_data$bathrooms)) +
    labs(title = "Price vs Bathroom Number")+
    labs(x = "Bathrooms") + labs(y = "Price")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```

```
## Warning: Removed 206 rows containing missing values (stat_boxplot).
```



YOUR WRITTEN ANSWER GOES HERE: Based on your plot, does the number of bathrooms affect the price? If so, how?

Based on the data having 3 bathrooms drives up the price of the unit. For the most part it doesn't matter too much how many bathrooms you have. As the number of bathrooms increased past 4.5 the price actually went down. I found this surprising however I confirmed this is the case by taking the sample of all apartments that had 6 bathrooms. Surprisingly there was only one, and that one was listed for only $800 in Berkeley.

Exercise 3

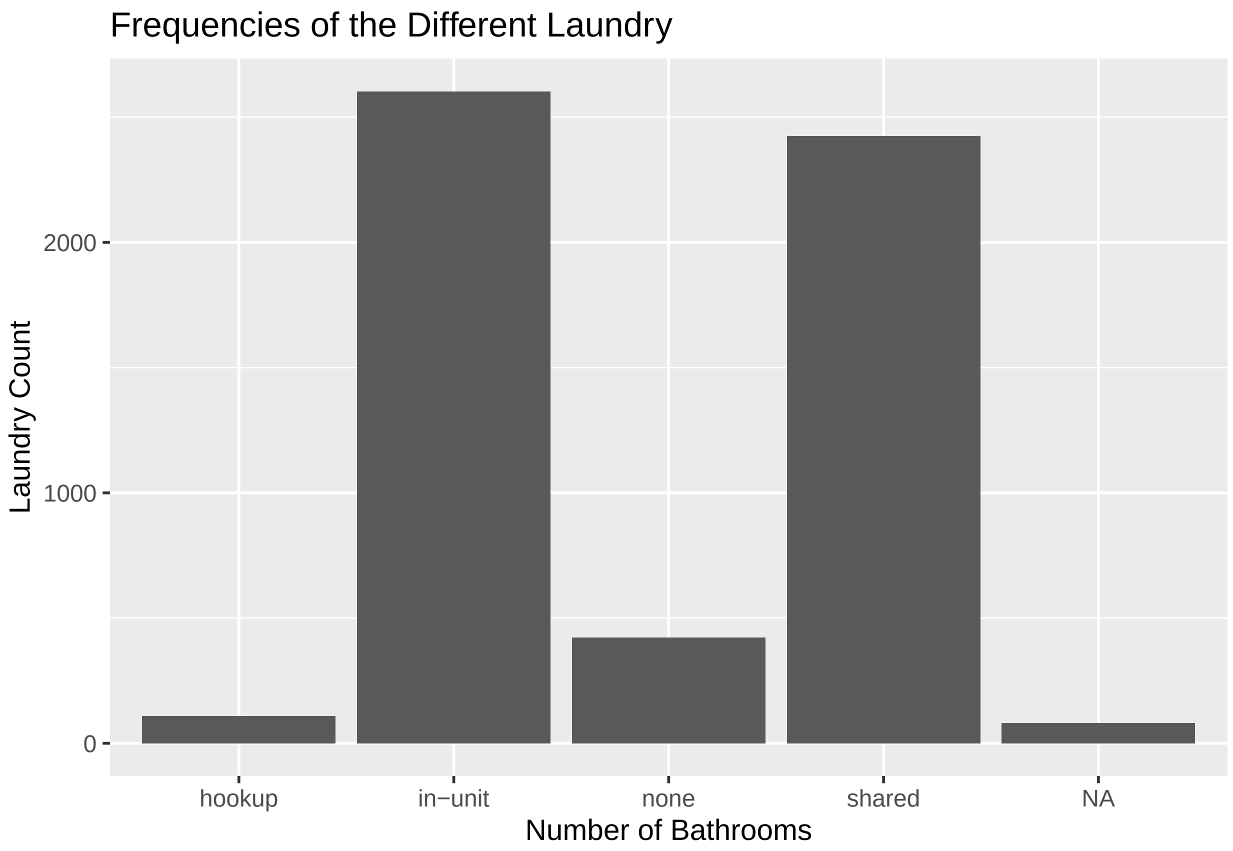
Bar plots are a good way to display the distribution (that is, the frequencies) of a 1-dimensional categorical feature. Groups of bars can be used to compare against a second categorical feature.

1. Get the subset of advertisements for apartments with 0 to 3 bedrooms.
2. For the subset, create a bar plot that shows the frequencies of the different laundry categories,

   a. grouped by the number of bedrooms (on the x-axis).
   b. Position the bars so that they are side-by-side rather than stacked.
   c. Add descriptive labels and a title to your plot.

Hint: You will need to set the `position` parameter for the geometry layer.

```
bath_comp <- subset(ba_data, ba_data$bathrooms <= 3)
ggplot(data = bath_comp, ) +
    geom_bar(aes((laundry), group = 1), position = position_dodge2()) +
    labs(title = "Frequencies of the Different Laundry", x = "Number of Bathrooms", y = "Laundry Count"
```



Exercise 4

The `table()` function computes the frequency of each unique value in a vector. If you want to inspect the distribution of categories in a categorical feature, the `table()` function is a quick alternative to a bar plot.

Use the `table()` function to compute a two-way table that shows the number of bedrooms versus the number of bathrooms for all of the advertisements in Oakland and San Francisco. Use the `dnn` parameter to label the appropriate dimensions of the table as "Bedrooms" and "Bathrooms".

Which combination of number of bedrooms and number of bathrooms is most common among apartment advertisements for these two cities?

```
sf_and_oak = subset(ba_data, ba_data$city == "San Francisco"| ba_data$city == "Oakland")
with(sf_and_oak, table(bedrooms, bathrooms), useNA = FALSE, dnn = "bedrooms",  "bathrooms")
```

```
##         bathrooms
## bedrooms   0   1 1.5   2 2.5   3 3.5   5
##        0   0 294   1   0   0   0   0   0
##        1   5 597   9   9   0   0   0   0
##        2   3 199   9 138   4   0   0   0
##        3   0  25   5  63   1   4   0   0
##        4   0   1   0  14   0   2   0   0
##        5   0   0   0   3   0   1   0   0
##        6   0   0   0   0   0   0   1   0
##        8   0   0   0   0   0   0   0   1
```

YOUR WRITTEN ANSWER GOES HERE: Which combination of number of bedrooms and number of bathrooms is most common among apartment advertisements for these two cities?

Based off of the table I found that the most common combination of bathroom and bedroom is having one of each.

Exercise 5

Given a vector, the `order()` function returns the vector of positions that sort the vector if used with the `[` operator. For example, suppose we want to order this vector from smallest to largest:

```
x = c(4, -6, 21, -2)
```

We can get the ordered vector by calling `order()` and then subsetting:

```
indices = order(x)
indices
```

```
## [1] 2 4 1 3
```

```
x[indices]
```

```
## [1] -6 -2  4 21
```

The order() function generalizes to data frames. If you get the ordering for a column, you can then use the ordering to subset the rows and reorder them.

Use order() to find the 5 least expensive apartment advertisements in Berkeley. Print their price, number of bedrooms, and title. Are all of them ads for apartments?

```
berk <- subset(ba_data, ba_data$city == 'Berkeley')
berk <- na.omit(berk)
cheap_berk <- order(berk$price)
cheap_berk <- cheap_berk[0:5]
r <- c(0,1,2,3,4)
for (i in r){
    print(berk$price[cheap_berk[i]])
    print(berk$bedrooms[cheap_berk[i]])
    print(berk$title[cheap_berk[i]])
}
```

```
## numeric(0)
## numeric(0)
## character(0)
## [1] 1500
## [1] 1
## [1] "$1500 / 1br - 500ft2 - Views, Charm, and History Renovated few mile to UC Berkeley (berkeley)"
## [1] 1500
## [1] 1
## [1] "$1500 / 1br - 500ft2 - Views, Charm, and History Renovated few mile to UC Berkeley (berkeley)"
## [1] 1525
## [1] 0
## [1] "$1525 / 600ft2 - Nice Studio Apt. - Convenient Location in Berkeley (berkeley)"
## [1] 1750
## [1] 1
## [1] "$1750 / 1br - 580ft2 - South Berkeley One Bedroom Apartment (berkeley)"
```

YOUR WRITTEN ANSWER GOES HERE: Are all of them ads for apartments?

Yes, they appear to all be ads for apartments. I'm not sure why the price is NA even after I omitted NAs.