Homework 2

Dakota Terry

Due Tuesday, February 7, 2022 at 5:00 PM in D2L

Instructions

Use this .Rmd file as a template for your homework. Please use D2L to turn in both the PDF output and your R Markdown (.Rmd) file. Your .Rmd file should compile on its own if it is downloaded by your instructor.

Load packages and data

Reminder: Packages that are not built-in to the original R installation need to first be installed in RStudio before the package will load into the current session. Refer to the homework instructions for package installation steps.

```
library(tidyverse)
library(dsbox)
```

Exercises

Exercise 1

The dataset has 13245 observations of 10 variables

```
dim(edibnb)
```

```
## [1] 13245    10
```

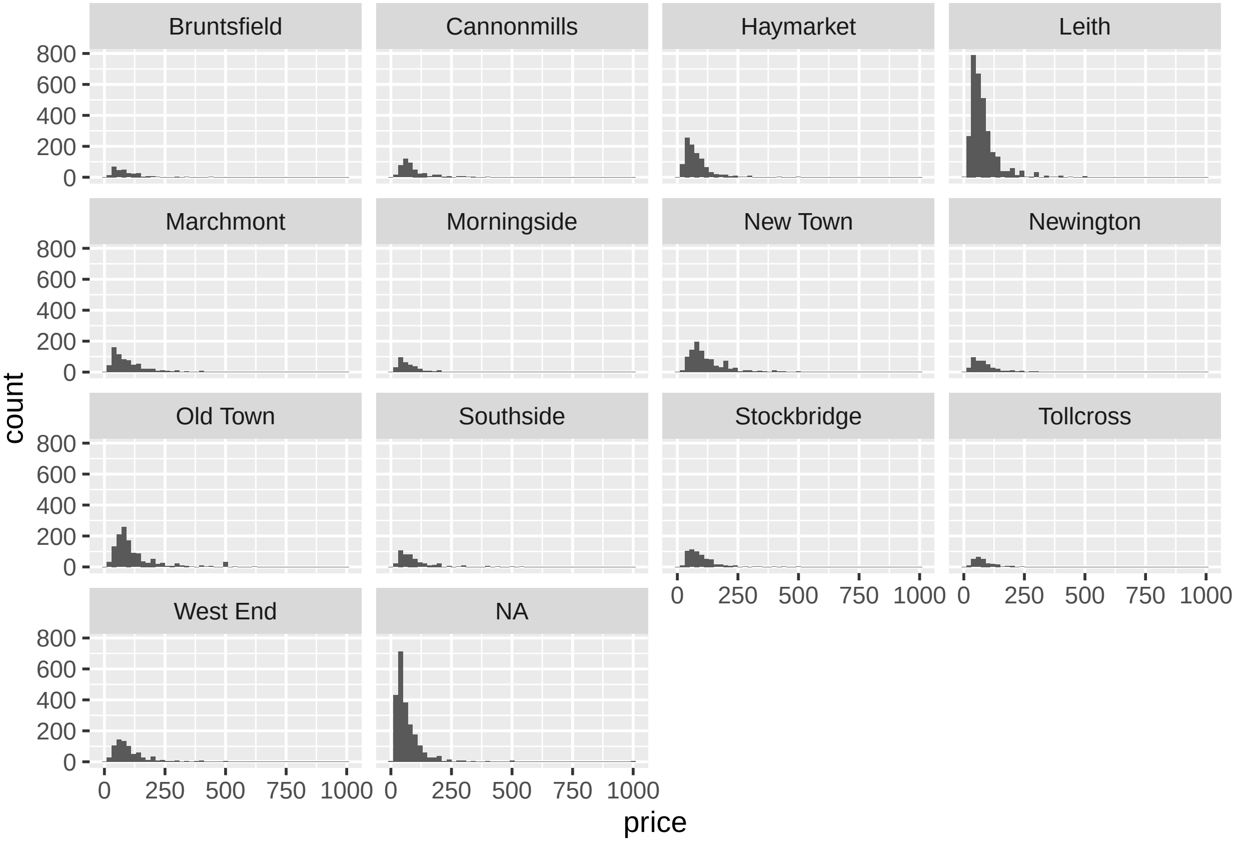
Exercise 2

Each row represents an airbnb that is listed for rental in Edinburgh, Scotland.

Exercise 3

I chose to wrap my facets into 4 rows and 4 columns. Ideally, i would have liked to lay them all out in a row for easy comparison, but that was too difficult with this many neighbourhoods to compare due to the small width of each facet. This way, you can see all of them side by side on the same scale and they all fit more comfortably on one screen. I chose a bin width of 20 because I felt that was the smallest value that showed a proper distribution curve without major volatility.

```
ggplot(data = edibnb, mapping = aes(x = price)) +
  geom_histogram(binwidth = 20) +
  facet_wrap(~neighbourhood , nrow = 4, ncol = 4)
```

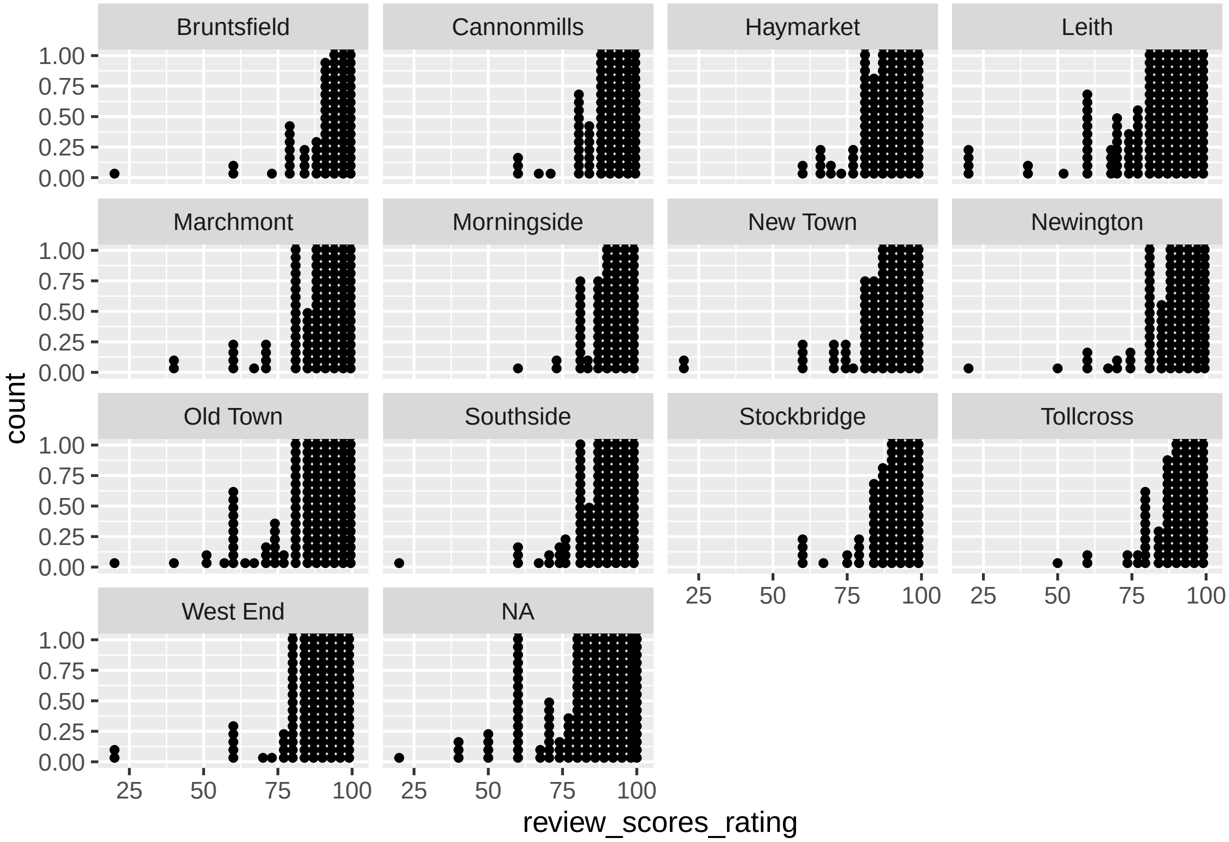


Exercise 4

I chose to make a facet-wrapped boxplot showing ratings by neighbourhood. Based on the charts, it is clear that most people give ratings of 100 to their airbnbs and very few people leave bad reviews. It is also evident that certain neighbourhoods get more mixed/bad reviews than others. The neighbourhoods of Cannonmills, Haymarket, Morningside, Stockbridge, and Tolicross had virtually no ratings below 50, and all other neighbourhoods had a certain, albeit small, proportion of very low ratings. One thing I noted was that Leith and the NA category had by far the most airbnb observations and also had the most variability in their ratings, which makes sense.

```
ggplot(data = edibnb, mapping = aes(x = review_scores_rating)) +
  geom_dotplot() +
  facet_wrap(~neighbourhood , nrow = 4, ncol = 4)
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with ‘binwidth‘.
```



Exercise 5

I found a free csv dataset online for work related injury claims for the years 2009-2018. To be honest, I'm not sure of the geographical location of these claims, as the dataset is not that great and does not break observations down by location. I used the dataset anyway, though, because I think it works for the purposes of this exercise.

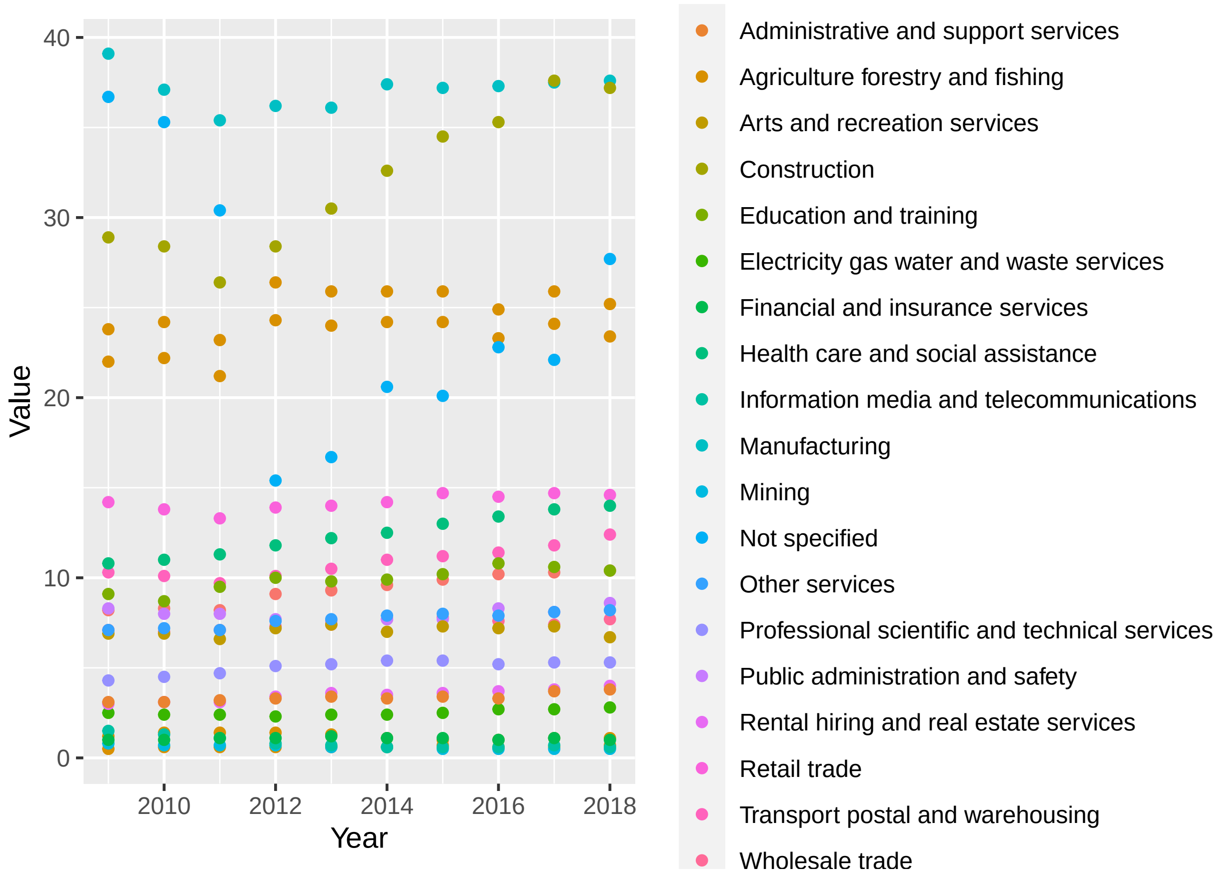
My first step was to remove the majority of observations where Industry was equal to "Total", as I didn't find this category useful… I also removed all observations in which Measure was not equal to "Number of claims in thousands" because there are observations in different units… this took me from nearly 2500 observations down to 230, but I found this final subset to be most useful for visualization.

For my visualization, I chose to create a scatter plot of the number of injury claims by year, and I color-coded my points based on the industry. My plot doesn't look great when you run and view it in this document because of size/scaling issues, but if you open the plot in a new window it scales down to the appropriate size and looks good. From this plot you can tell which industries have the most work-related injuries (Mining was the worst, followed by construction and then forestry/fishing). It was also interesting to me that it appears most industries have had an increasing rate of injury over time (except mining), and it seems to me like work should be getting safer with all the new technology we have, but it appears to be getting more dangerous. This could be explained by the fact that people are more apt to file claims in recent years than they used to be.

```
injuries <- read.csv("injury-statistics-work-related-claims-2018-csv.csv")
```

```r
injuries = subset(injuries, Industry!="Total" )
injuries = subset(injuries, Measure == "Number of claims in thousands")

 ggplot(data = injuries, mapping = aes(y=Value,x=Year, color = Industry)) +
  geom_point()
```



Cite Sources

https://ggplot2.tidyverse.org/reference/index.html

https://www.stats.govt.nz/large-datasets/csv-files-for-download/