HW 02 - Airbnb listings in Edinburgh

Tina Huynh

Figure 1: Photo by Madeleine Kohler on Unsplash

Once upon a time, people traveled all over the world, and some stayed in hotels and others chose to stay in other people's houses that they booked through Airbnb. Recent developments in Edinburgh regarding the growth of Airbnb and its impact on the housing market means a better understanding of the Airbnb listings is needed. Using data provided by Airbnb, we can explore how Airbnb availability and prices vary by neighbourhood.

Getting started

```
**IMPORTANT:** If there is no GitHub repo created for you for this assignment, it means I didn't have y
```

Go to the course GitHub organization and locate your homework repo, which should be named `hw-02-airbnb-edi-YOUR_GITHUB_USERNAME`. Grab the URL of the repo, and clone it in RStudio. First, open the R Markdown document `hw-02.Rmd` and Knit it. Make sure it compiles without errors. The output will be in the file markdown `.md` file with the same name.

Warm up

Before we introduce the data, let's warm up with some simple exercises.

- Update the YAML, changing the author name to your name, and **knit** the document.
- Commit your changes with a meaningful commit message.
- Push your changes to GitHub.
- Go to your repo on GitHub and confirm that your changes are visible in your Rmd **and** md files. If anything is missing, commit and push again.

Packages

We'll use the **tidyverse** package for much of the data wrangling and visualization and the data lives in the **dsbox** package. These packages are already installed for you. You can load them by running the following in your Console:

```
install.packages("devtools")
devtools::install_github("tidyverse/dsbox")
library(tidyverse)
library(dsbox)
```

Data

The data can be found in the **dsbox** package, and it's called `edibnb`. Since the dataset is distributed with the package, we don't need to load it separately; it becomes available to us when we load the package.

You can view the dataset as a spreadsheet using the `View()` function. Note that you should not put this function in your R Markdown document, but instead type it directly in the Console, as it pops open a new window (and the concept of popping open a window in a static document doesn't really make sense…). When you run this in the console, you'll see the following **data viewer** window pop up.

```
View(edibnb)
```

You can find out more about the dataset by inspecting its documentation, which you can access by running `?edibnb` in the Console or using the Help menu in RStudio to search for `edibnb`.

Exercises

```
**Hint:** The Markdown Quick Reference sheet has an example of inline R code that might be helpful. You
```

1. How many observations (rows) does the dataset have? Instead of hard coding the number in your answer, use inline code. `13,245 entities in total`
2. Run `View(edibnb)` in your Console to view the data in the data viewer. What does each row in the dataset represent?
   Each data row represents a listing for a Airbnb rental

Knit, commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.

Each column represents a variable. We can get a list of the variables in the data frame using the `names()` function.

```
names(edibnb)
```

You can find descriptions of each of the variables in the help file for the dataset, which you can access by running `?edibnb` in your Console.

```
**Note:** The plot will give a warning about some observations with non-finite values for price being r
```

3. Create a faceted histogram where each facet represents a neighbourhood and displays the distribution of Airbnb prices in that neighbourhood. Think critically about whether it makes more sense to stack the facets on top of each other in a column, lay them out in a row, or wrap them around. Along with your visualization, include your reasoning for the layout you chose for your facets.

```
ggplot(edibnb, aes(x = price)) +
  geom_histogram(aes(y = after_stat(density)), bins = 25, boundary = 10, closed = "left") +
  coord_cartesian(xlim = c(0, quantile(edibnb$price, 0.99, na.rm = TRUE))) +
  facet_wrap(~ neighbourhood, ncol = 5) +
  labs(
    title = "Edinburgh Airbnb Price Distributions by neighbourhood",
    subtitle = "Density-scaled histograms; x-axis capped at the 99th percentile",
    x = "Price (GBP)",
    y = "Density",
    caption = "Source: dsbox::edibnb (Inside Airbnb / Kaggle)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    strip.text = element_text(size = 10),
    panel.grid.minor = element_blank(),
    axis.text = element_text(size = 6),
    plot.margin = unit(c(0.5, 0.25, 0.5, 0.25), "cm")
  )
```

Let's de-construct this code:

- `ggplot()` is the function we are using to build our plot, in layers.
- In the first layer we always define the data frame as the first argument. Then, we define the mappings between the variables in the dataset and the **aes**thetics of the plot (e.g. x and y coordinates, colours, etc.).
- In the next layer we represent the data with **geom**etric shapes, in this case with a histogram. You should decide what makes a reasonable bin width for the histogram by trying out a few options.
- In the final layer we facet the data by neighbourhood.

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.*

4. Use a single pipeline to identity the neighbourhoods with the top five median listing prices. Then, in another pipeline filter the data for these five neighbourhoods and make ridge plots of the distributions of listing prices in these five neighbourhoods. In a third pipeline calculate the minimum, mean, median, standard deviation, IQR, and maximum listing price in each of these neighbourhoods. Use the visualization and the summary statistics to describe the distribution of listing prices in the neighbourhoods. (Your answer will include three pipelines, one of which ends in a visualization, and a narrative.)

```r
top5_neighbourhoods <- edibnb %>%
  filter(!is.na(price), price > 0, !is.na(neighbourhood)) %>%
  group_by(neighbourhood) %>%
  summarize(median_price = median(price, na.rm = TRUE)) %>%
  arrange(desc(median_price)) %>%
  slice_head(n = 5)

top5_neighbourhoods
```

```r
library(ggridges)

edibnb %>%
  filter(!is.na(price), price > 0, neighbourhood %in% top5_neighbourhoods$neighbourhood) %>%
  ggplot(aes(x = price, y = neighbourhood, fill = neighbourhood)) +
  geom_density_ridges(alpha = 0.6, scale = 1.2, rel_min_height = 0.01) +
  coord_cartesian(xlim = c(0, quantile(dsbox::edibnb$price, 0.99, na.rm = TRUE))) +
  labs(
    title = "Price Distributions of Top 5 Most Expensive neighbourhoods in Edinburgh",
    x = "Listing Price (GBP)",
    y = "neighbourhood"
  ) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none")
```

```r
summary_stats <- edibnb %>%
  filter(!is.na(price), price > 0, neighbourhood %in% top5_neighbourhoods$neighbourhood) %>%
  group_by(neighbourhood) %>%
  summarize(
    min_price = min(price, na.rm = TRUE),
    mean_price = mean(price, na.rm = TRUE),
    median_price = median(price, na.rm = TRUE),
    sd_price = sd(price, na.rm = TRUE),
    iqr_price = IQR(price, na.rm = TRUE),
    max_price = max(price, na.rm = TRUE)
  )

summary_stats
```

4. Create a visualization that will help you compare the distribution of review scores (`review_scores_rating`) across neighbourhoods. You get to decide what type of visualization to create and there is more than one correct answer! In your answer, include a brief interpretation of how Airbnb guests rate properties in general and how the neighbourhoods compare to each other in terms of their ratings.

```r
ggplot(edibnb %>%
  filter(!is.na(review_scores_rating), !is.na(neighbourhood)) %>%
  mutate(
    neighbourhood = fct_reorder(neighbourhood, review_scores_rating, .fun = median, .na_rm = TRUE)
  ), aes(x = neighbourhood, y = review_scores_rating)) +
  geom_boxplot(fill = "steelblue", alpha = 0.6, outlier.alpha = 0.3) +
  coord_flip() +
  labs(
    title = "Distribution of Airbnb Review Scores by neighbourhood in Edinburgh",
```

```
    x = "neighbourhood",
    y = "Review Score (0-100)"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    axis.text.y = element_text(size = 9)
  )
```

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards and review the md document on GitHub to make sure you're happy with the final state of your work.*