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? 13245 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 an Airbnb rental in Edinburgh.

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.

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

```
names(edibnb)
```

```
##  [1] "id"                  "price"               "neighbourhood"
##  [4] "accommodates"        "bathrooms"           "bedrooms"
##  [7] "beds"                "review_scores_rating" "number_of_reviews"
## [10] "listing_url"
```

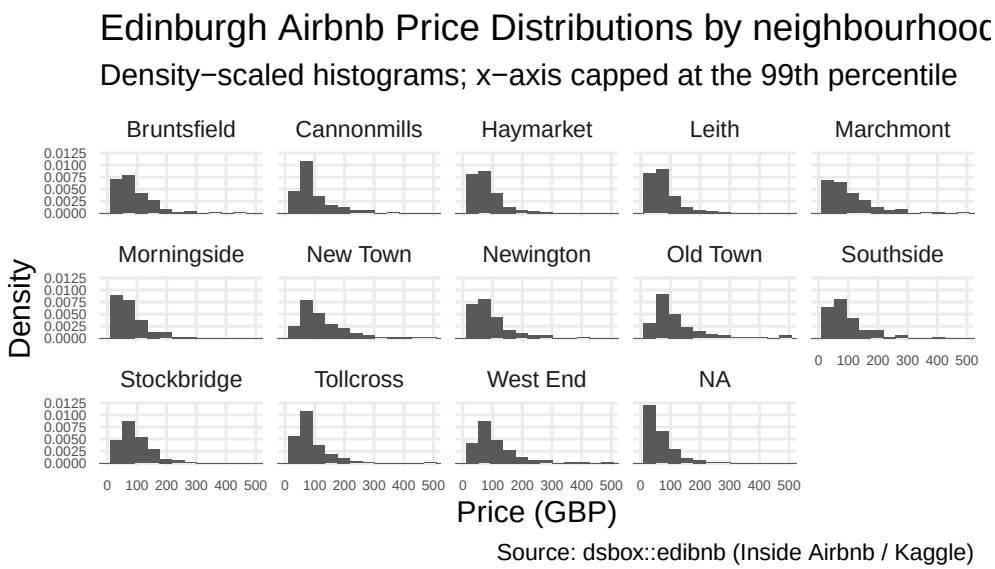
You can find descriptions of each of the variables in the help file for the dataset, which you can access by running `?edibnb` in your Console.

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

1. 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")
  )
```

```
## Warning: Removed 199 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



I chose to use `facet_wrap()` with 5 columns because:

1. Edinburgh has multiple neighborhoods, and wrapping them allows for better use of the horizontal space than a single column or row. This way, we can see more neighborhoods at once without excessive scrolling.
2. With 5 columns, each facet is still large enough to see the distribution clearly and not too compressed.
3. This layout makes it easier to compare neighborhoods at a glance than stacking them in a column. Therefore, it is more efficient for visual comparison.
4. The wrapped layout provides a good balance between detail and overall comparison view because it avoids excessive white space that would occur with too few columns.

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.*

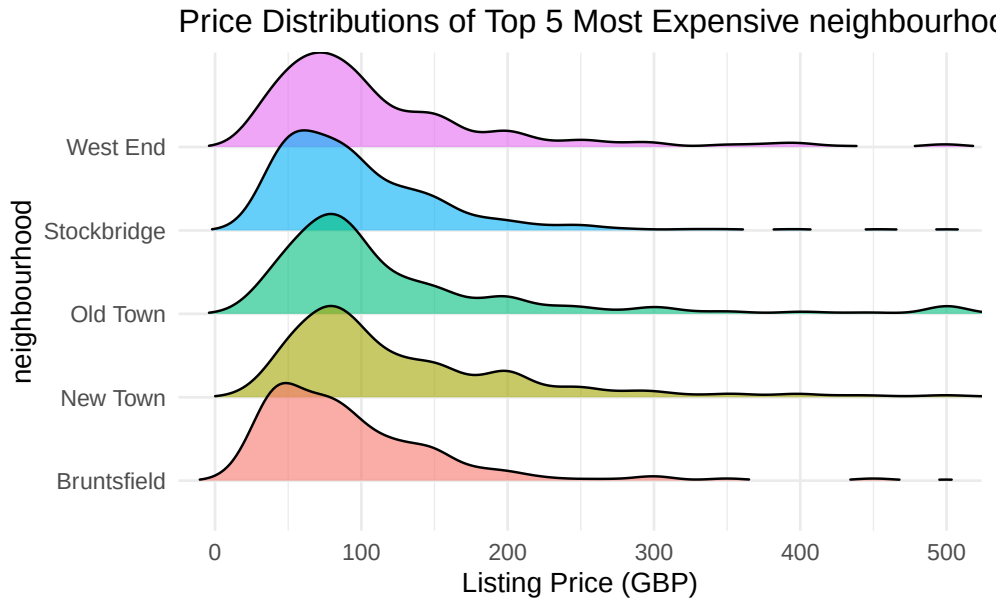
4. Use a single pipeline to identity the neighbourhoods with the top five median listing prices. Then, in another pipeline filter the data for these five neighbourhoods and make ridge plots of the distributions of listing prices in these five neighbourhoods. In a third pipeline calculate the minimum, 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.)

```
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
```

```
## # A tibble: 5 x 2
##   neighbourhood median_price
##   <chr>                <dbl>
## 1 New Town               100
## 2 Old Town                90
## 3 West End                90
## 4 Stockbridge             85
## 5 Bruntsfield             80
```

```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")
```

```
## Picking joint bandwidth of 13.8
```




```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
```

```
## # A tibble: 5 x 7
##   neighbourhood min_price mean_price median_price sd_price iqr_price max_price
##   <chr>             <dbl>      <dbl>        <dbl>    <dbl>     <dbl>     <dbl>
## 1 Bruntsfield          10       99.4           80     90.2      72.5       900
## 2 New Town             12      136.           100    109.       86.5       999
## 3 Old Town             15      128.            90    110.       76         999
## 4 Stockbridge          21      104.            85     77.6      66         750
## 5 West End             19      116.            90     93.3      80         999
```

The price distributions in all five neighborhoods demonstrate pronounced positive skewness, characterized by a concentration of listings at moderate price points with an extended tail toward higher values. This asymmetry suggests a market structure where most properties maintain accessible pricing while a select subset commands premium rates.

New Town, with the highest median price (£100), exhibits the most pronounced variance (standard deviation of £108.59), indicating considerable price stratification within this historically significant district. This variability likely reflects the neighborhood's diverse property portfolio, ranging from modest apartments to luxury Georgian residences. The substantial interquartile range (£86.50) further confirms this price diversity, suggesting potential investment opportunities across various market segments.

Stockbridge and West End complete the top five neighborhoods, displaying intermediate pricing patterns. Notably, Stockbridge exhibits the lowest standard deviation (£77.57) among these premium areas, potentially signifying a more uniform housing stock. The median price (£85) and interquartile range (£76) suggest a balanced market, appealing to both mid-range and higher-end renters.

Old Town, despite its status as a primary tourist destination, presents a slightly lower median price (£90) compared to New Town, while maintaining comparable variability (standard deviation of £109.57). This pattern may reflect the neighborhood's mixed accommodation offerings, from heritage buildings to more contemporary developments. The area's tourism appeal evidently supports elevated price points, with maximum values reaching £999.

Bruntsfield demonstrates a more concentrated price distribution (standard deviation of £90.17) relative to its median (£80), suggesting greater homogeneity in property types and quality within this neighborhood. This characteristic could indicate a more stable, predictable market with potentially lower investment risk.

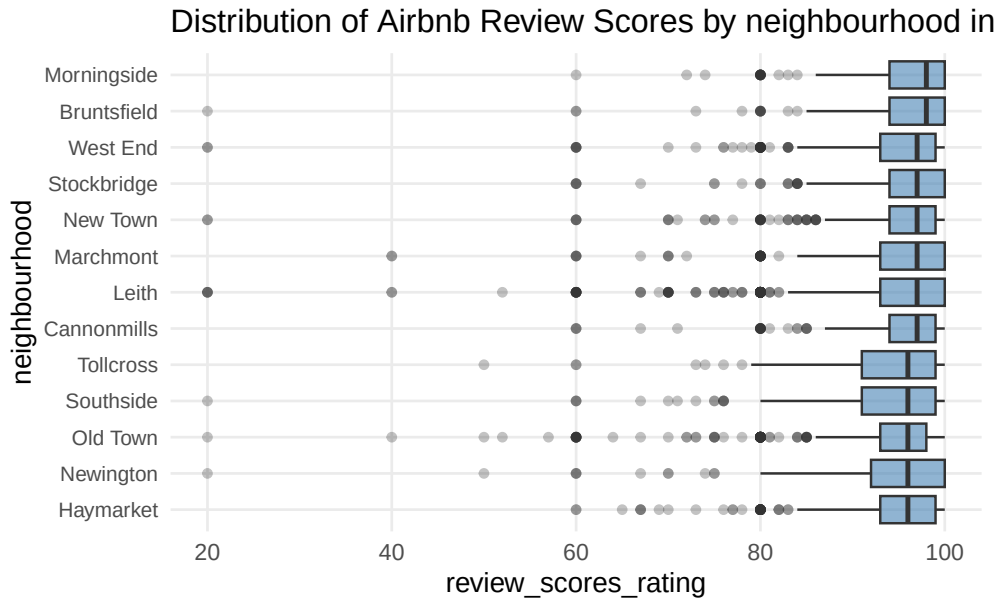
The minimum listing prices remain relatively consistent across all five neighborhoods (ranging from £10-£21), indicating that even in Edinburgh's most exclusive districts, budget accommodation options exist. This accessibility at the lower end contrasts sharply with the maximum values, which reach £999 in multiple neighborhoods, highlighting significant market segmentation.

These pricing patterns reflect Edinburgh's complex residential landscape, where historical significance, architectural heritage, and proximity to cultural attractions interact to create distinctive neighborhood value propositions. The data suggests that New Town and Old Town command premium pricing due to their central location and historical importance, while areas like Bruntsfield may offer more consistent pricing expectations for both investors and visitors seeking accommodation.

1. 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.

```
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)
  )
```



The visualization of review score distributions reveals nuanced patterns in guest satisfaction across Edinburgh's diverse neighborhoods. Examination of the boxplots indicates a predominantly positive guest experience throughout the city, with most neighborhoods achieving median ratings exceeding 90 on the 100-point scale. This high baseline suggests Edinburgh's Airbnb market generally delivers satisfactory accommodations regardless of location.

Notable in this analysis is the inverse relationship between accommodation cost and guest satisfaction. The neighborhoods commanding premium pricing—particularly New Town and West End—paradoxically rank lower in guest satisfaction than more moderately priced areas. This phenomenon may reflect guests' value expectations; higher-priced accommodations potentially trigger heightened expectations that prove challenging to fulfill consistently. Conversely, neighborhoods like Stockbridge, South Side, and Leith, which offer more moderate pricing, achieve superior satisfaction ratings, possibly because they exceed expectations relative to their cost.

The remarkably compressed interquartile ranges observed across most neighborhoods indicate consistency in service quality within each district. This homogeneity suggests established hosting standards within neighborhood communities or similar property types dominating specific areas. The narrow distribution bands also imply reliability in guest experiences, a valuable attribute for travelers selecting accommodations.

While outliers appear across all neighborhoods, their relatively sparse distribution compared to the concentration of positive ratings indicates isolated instances of substandard experiences rather than systemic quality issues. These outliers likely represent properties with specific deficiencies or management problems rather than reflecting neighborhood-wide concerns.

The ranking pattern observed challenges conventional assumptions about accommodation quality being directly proportional to price or centrality. Neighborhoods traditionally considered less prestigious outperform Edinburgh's most renowned districts in guest satisfaction, suggesting that intangible factors—host responsiveness, accuracy of listing descriptions, cleanliness standards, or value perception—may influence ratings more significantly than location prestige or property luxury.

For potential visitors, this analysis provides evidence-based guidance suggesting that excellent accommodation experiences are available throughout Edinburgh, with some less centrally located or lower-priced neighborhoods potentially offering superior guest experiences compared to their more expensive counterparts. For hosts and property managers, these patterns highlight the importance of expectation management and service quality in achieving positive reviews, particularly in high-priced areas where guest expectations may be correspondingly elevated.