Assignment1

2025-10-02

Question: How do sunshine hours influence Airbnb occupancy rates across different US cities from January to August?
- Using the provided sunshine hours dataset and the additional Airbnb occupancy data from https://www.listingok.com/en/airbnb-occupancy/united-states/ for the cities Seattle, Phoenix, New York, Chicago, Miami, and Houston, we can explore how sunshine hours influence the Airbnb occupancies.
- By joining the two datasets on city and month, each observation contains both the average monthly sunshine hours and the corresponding Airbnb occupancy rate.This allows to investigate the question: “How do monthly sunshine hours influence Airbnb occupancy rates across these cities from January to August?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
sunshine_data <- read_csv("/Users/sowjanyapadala/Desktop/Coursework/Q4/Data_Visualization/Assignments/su
```

```
## Rows: 84 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): city, month
## dbl (4): sunshine, lat, lon, monthnum
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
occupancy_data <- read_csv("/Users/sowjanyapadala/Desktop/Coursework/Q4/Data_Visualization/Assignments/A
```

```
## Rows: 48 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): city, month
## dbl (1): OccupancyRate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(sunshine_data)
```

```
## # A tibble: 6 x 6
##   sunshine city      lat   lon month monthnum
##      <dbl> <chr>    <dbl> <dbl> <chr>    <dbl>
## 1       69 Seattle  47.6 -122. Jan          0
## 2      108 Seattle  47.6 -122. Feb          1
## 3      178 Seattle  47.6 -122. Mar          2
## 4      207 Seattle  47.6 -122. Apr          3
## 5      253 Seattle  47.6 -122. May          4
## 6      268 Seattle  47.6 -122. Jun          5
```

```
head(occupancy_data)
```

```
## # A tibble: 6 x 3
##   city    month OccupancyRate
##   <chr>   <chr>         <dbl>
## 1 Seattle Jan            48.6
## 2 Seattle Feb            56.3
## 3 Seattle Mar            54.9
## 4 Seattle Apr            51.8
## 5 Seattle May            58.7
## 6 Seattle Jun            71.2
```

```
cat("Cities in the sunshine data:", unique(sunshine_data$city), "\n")
```

```
## Cities in the sunshine data: Seattle Phoenix New York Chicago Houston Miami Salt Lake City
```

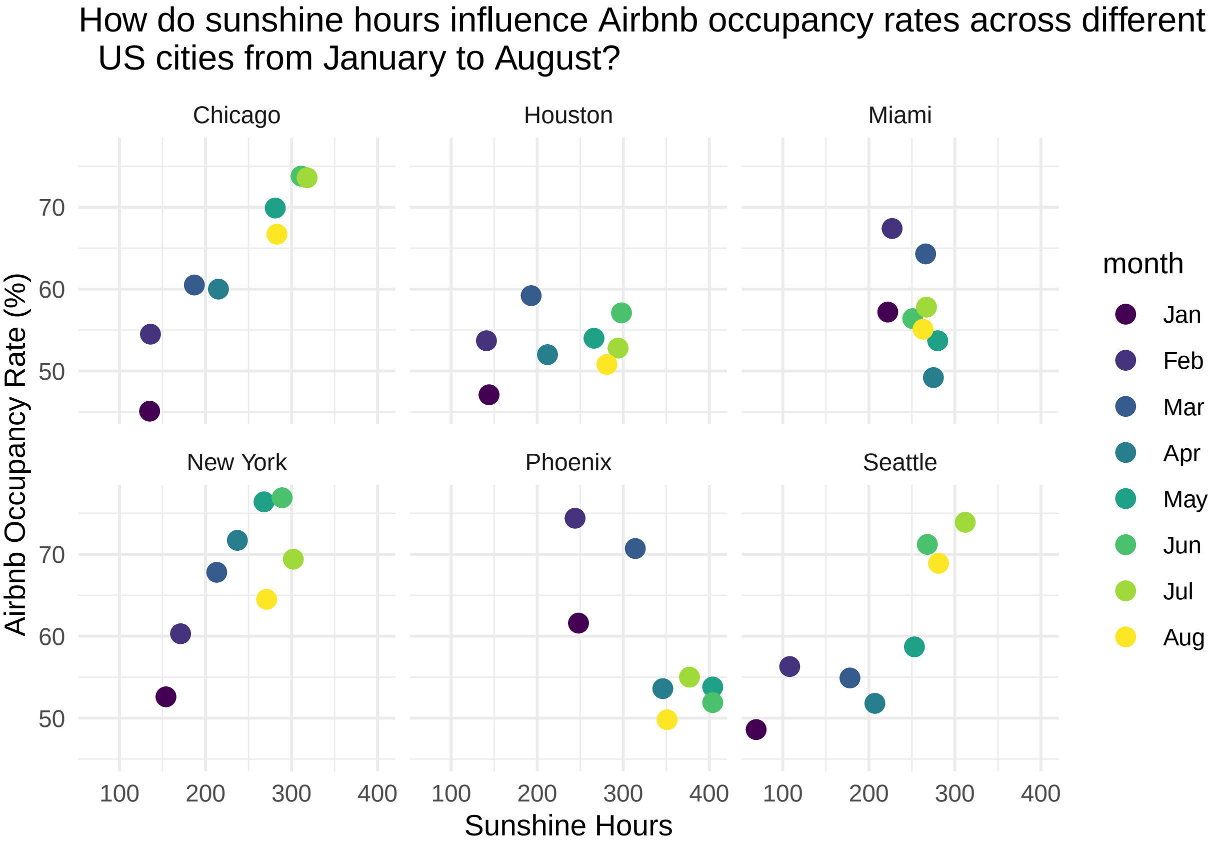
```
# Print unique cities in Airbnb occupancy data
cat("Cities in the Airbnb Occupancy data:", unique(occupancy_data$city), "\n")
```

```
## Cities in the Airbnb Occupancy data: Seattle Phoenix New York Chicago Houston Miami
```

```
merged_df <- occupancy_data %>% inner_join(sunshine_data, by=c("city", "month"))

#month levels
month_levels <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug")

# Convert Month to factor
merged_df$month <- factor(merged_df$month, levels = month_levels, ordered = TRUE)

#arrange by City and Month
merged_df <- merged_df %>%
  arrange(city, month)
merged_df
```

```
## # A tibble: 48 x 7
##    city    month OccupancyRate sunshine   lat   lon monthnum
##    <chr>   <ord>         <dbl>    <dbl> <dbl> <dbl>    <dbl>
##  1 Chicago Jan            45.1      135  41.9 -87.6        0
##  2 Chicago Feb            54.5      136  41.9 -87.6        1
##  3 Chicago Mar            60.5      187  41.9 -87.6        2
##  4 Chicago Apr            60        215  41.9 -87.6        3
##  5 Chicago May            69.9      281  41.9 -87.6        4
##  6 Chicago Jun            73.8      311  41.9 -87.6        5
##  7 Chicago Jul            73.6      318  41.9 -87.6        6
##  8 Chicago Aug            66.7      283  41.9 -87.6        7
##  9 Houston Jan            47.1      144  29.8 -95.4        0
## 10 Houston Feb            53.7      141  29.8 -95.4        1
## # i 38 more rows
```

```
library(ggplot2)

ggplot(merged_df, aes(x = sunshine, y = OccupancyRate)) +
  geom_point(aes(color = month), size = 3) +          # points colored by month
  facet_wrap(~city) +                                # separate plot per city
  labs(title = "How do sunshine hours influence Airbnb occupancy rates across different
  US cities from January to August?",
       x = "Sunshine Hours",
       y = "Airbnb Occupancy Rate (%)") +
  theme_minimal()
```



The above faceted scatter plot visualizes the relationship between monthly sunshine hours and Airbnb occupancy rates from January to August across six U.S. cities—Seattle, Phoenix, New York, Chicago, Miami, and Houston. The motivation for choosing faceted scatter plot for this analysis is because it allows us to examine the relationship between sunshine hours and Airbnb occupancy rates from January to August separately for each U.S. city while still maintaining a consistent visual framework for comparison. If we plotted all cities together in a single scatter plot, overlapping data points could obscure meaningful patterns. By faceting that is, creating small multiples for each city we can isolate the local relationship within each geographic context while keeping the same x- and y-scales across all panels. The data layer was constructed by integrating two datasets from the sunshine dataset, the variables city name, month, and average monthly sunshine hours were selected from the Airbnb dataset, the variable average monthly occupancy rate was extracted. These were joined on the variables city and month, producing a unified dataset.

-Along with the data, we need a specification of which variables are mapped to which aesthetics following this the plot maps sunshine hours to the x-position, Airbnb occupancy rate to the y-position, and month to color. The geometric object (geom) is point, which represents each observation individually without aggregation, and the statistical transformation (stat) used is the default, meaning the raw data are displayed without modification.The faceting specification (facet_wrap(~city)) divides the data by city, creating small multiples that isolate local relationships yet share identical scales.

-These layered design decisions collectively ensure that the plot clearly reveals how the relationship between sunshine hours and occupancy differs across regions. For example, in Seattle and Chicago, we can clearly observe a strong positive association as sunshine hours increase from winter(jan) to summer(August), occupancy rates rise sharply, indicating that there are more booking during sunnier months. In contrast, in Miami, the pattern is quite different the occupancy rate is relatively higher during the winter months compared to the peak summer period. This occurs probably because Miami's sunshine hours remain consistently high throughout the year, and its winter season attracts visitors escaping colder northern climates.