Gandre_Final_project

Paritosh Gandre

2023-11-23

fontfamily: mathpazo fontsize: 11pt header-includes: - urlcolor: blue —

libraries:

```
## Warning: package 'cowplot' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x lubridate::stamp() masks cowplot::stamp()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
## Loading required package: car
##
## Loading required package: carData
##
##
## Attaching package: 'car'
##
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
##
## The following object is masked from 'package:purrr':
##
##     some
##
##
## Loading required package: effects
##
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
## Warning: package 'viridis' was built under R version 4.3.2
```

```
## Loading required package: viridisLite
```

Dataset:

```
NY_data = readr::read_csv("D:/DATA_ANALYTICS/NEW YORK AIR BNB/AB_NYC_2019.csv")
```

```
## Rows: 48895 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (5): name, host_name, neighbourhood_group, neighbourhood, room_type
## dbl  (10): id, host_id, latitude, longitude, price, minimum_nights, number_o...
## date  (1): last_review
##
## 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(NY_data)
```

```
## # A tibble: 6 x 16
##      id name        host_id host_name neighbourhood_group neighbourhood latitude
##   <dbl> <chr>         <dbl> <chr>     <chr>               <chr>            <dbl>
## 1  2539 Clean & qu~    2787 John      Brooklyn            Kensington        40.6
## 2  2595 Skylit Mid~    2845 Jennifer  Manhattan           Midtown           40.8
## 3  3647 THE VILLAG~    4632 Elisabeth Manhattan           Harlem            40.8
## 4  3831 Cozy Entir~    4869 LisaRoxa~ Brooklyn            Clinton Hill      40.7
## 5  5022 Entire Apt~    7192 Laura     Manhattan           East Harlem       40.8
## 6  5099 Large Cozy~    7322 Chris     Manhattan           Murray Hill       40.7
## # i 9 more variables: longitude <dbl>, room_type <chr>, price <dbl>,
## #   minimum_nights <dbl>, number_of_reviews <dbl>, last_review <date>,
## #   reviews_per_month <dbl>, calculated_host_listings_count <dbl>,
## #   availability_365 <dbl>
```

```
ny_clean = NY_data |> na.omit()
```

Introduction [15 points]

- What research question(s) would you like to answer?

1 Distribution of price in different neighbourhood

```
library(ggplot2)

# neighbourhood
top_neighbourhoods = ny_clean |>
  group_by(neighbourhood) |>
  summarize(Private = mean(price[room_type == "Private room"]),
            Entire_Home = mean(price[room_type == "Entire home/apt"]),
            Shared_room = mean(price[room_type == "Shared room"])) |>
  top_n(10, Private) |>
  arrange(desc(Private))
top_neighbourhoods
```

```
## # A tibble: 10 x 4
##    neighbourhood     Private Entire_Home Shared_room
```

```
##    <chr>               <dbl>       <dbl>       <dbl>
##  1 Midtown              222.        290.        98.6
##  2 Breezy Point         195         NaN        NaN
##  3 West Village         180.        262.       180
##  4 Belle Harbor         178.        148.       NaN
##  5 Theater District     174.        303.        79
##  6 Coney Island         160         159.        32.2
##  7 Holliswood           159         NaN        NaN
##  8 Flatiron District    156.        323.       NaN
##  9 DUMBO                139         206.       NaN
## 10 Lower East Side      138.        211.        52.8
```

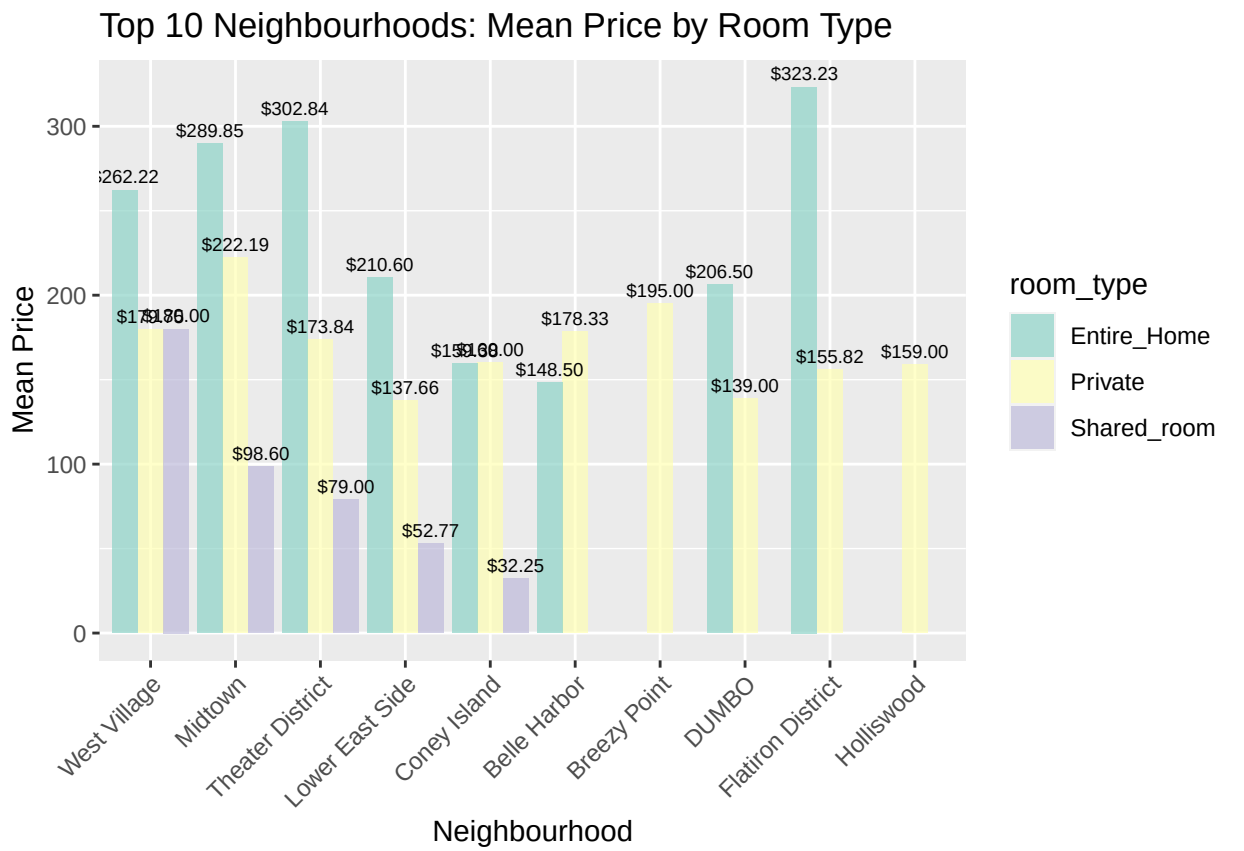
```
top_neighbourhoods_long =top_neighbourhoods |>  pivot_longer(-neighbourhood,names_to = "room_type",valu
top_neighbourhoods_long
```

```
## # A tibble: 30 x 3
##    neighbourhood room_type   mean_price
##    <chr>         <chr>            <dbl>
##  1 Midtown       Private          222.
##  2 Midtown       Entire_Home      290.
##  3 Midtown       Shared_room       98.6
##  4 Breezy Point  Private          195
##  5 Breezy Point  Entire_Home      NaN
##  6 Breezy Point  Shared_room      NaN
##  7 West Village  Private          180.
##  8 West Village  Entire_Home      262.
##  9 West Village  Shared_room      180
## 10 Belle Harbor  Private          178.
## # i 20 more rows
```

```
ggplot(top_neighbourhoods_long, aes(x = reorder(neighbourhood, -mean_price), y = mean_price, fill = roo
  geom_bar(stat = "identity", position = "dodge", alpha = 0.7) +
  geom_text(aes(label = sprintf("$%.2f", mean_price)), position = position_dodge(width = 0.9), vjust = 
  labs(title = "Top 10 Neighbourhoods: Mean Price by Room Type",x = "Neighbourhood",y = "Mean Price") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")
```

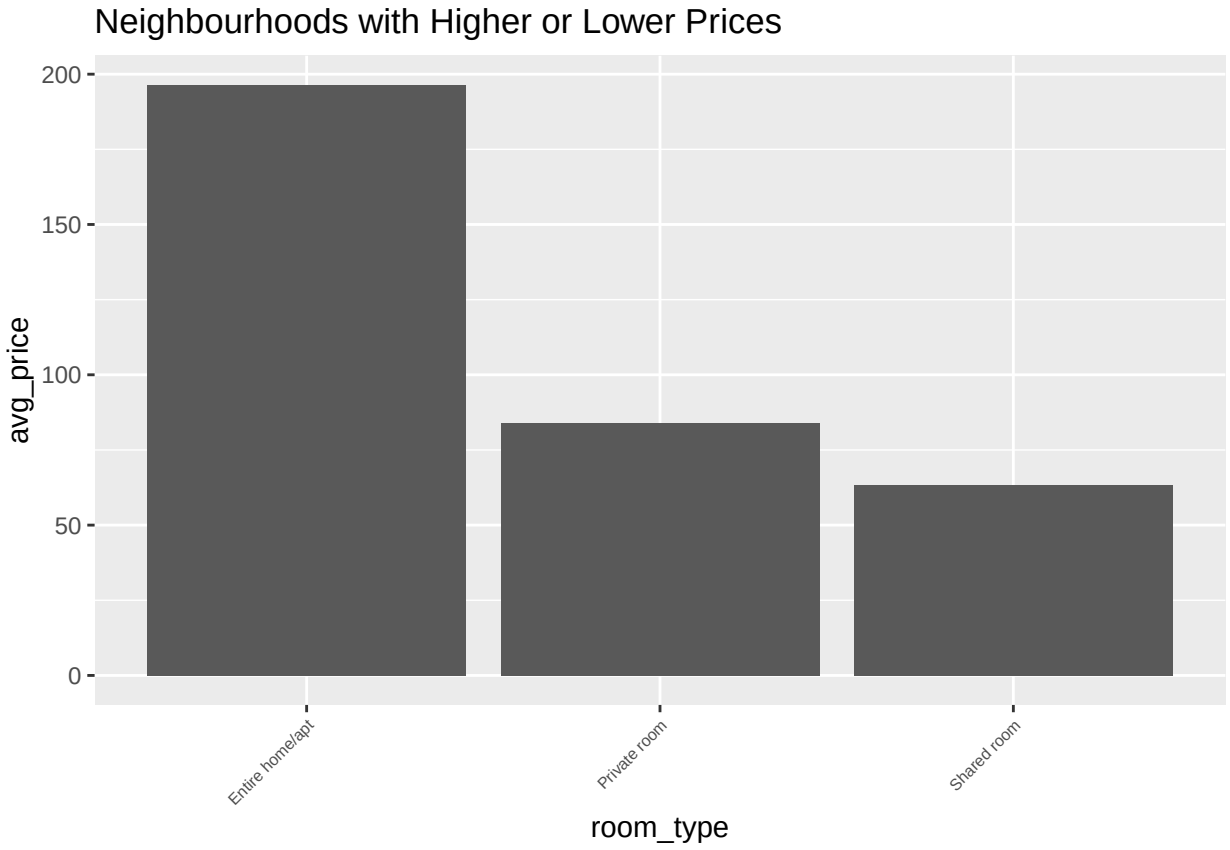
```
## Warning: Removed 7 rows containing missing values (`geom_bar()`).
```

```
## Warning: Removed 7 rows containing missing values (`geom_text()`).
```



2 How does room type affect price?

```
ny_clean |>
  group_by(room_type)  |>
  summarize(avg_price = mean(price)) |>
  ggplot(aes(room_type,avg_price)) +
  geom_bar(stat = "identity") +
  labs(title = "Neighbourhoods with Higher or Lower Prices")+
  theme(axis.text.x = element_text(angle =45, hjust = 1,size = 6))
```



3 Neighbourhood with higher or lower price

4 Correlation between number of reviews and pricing

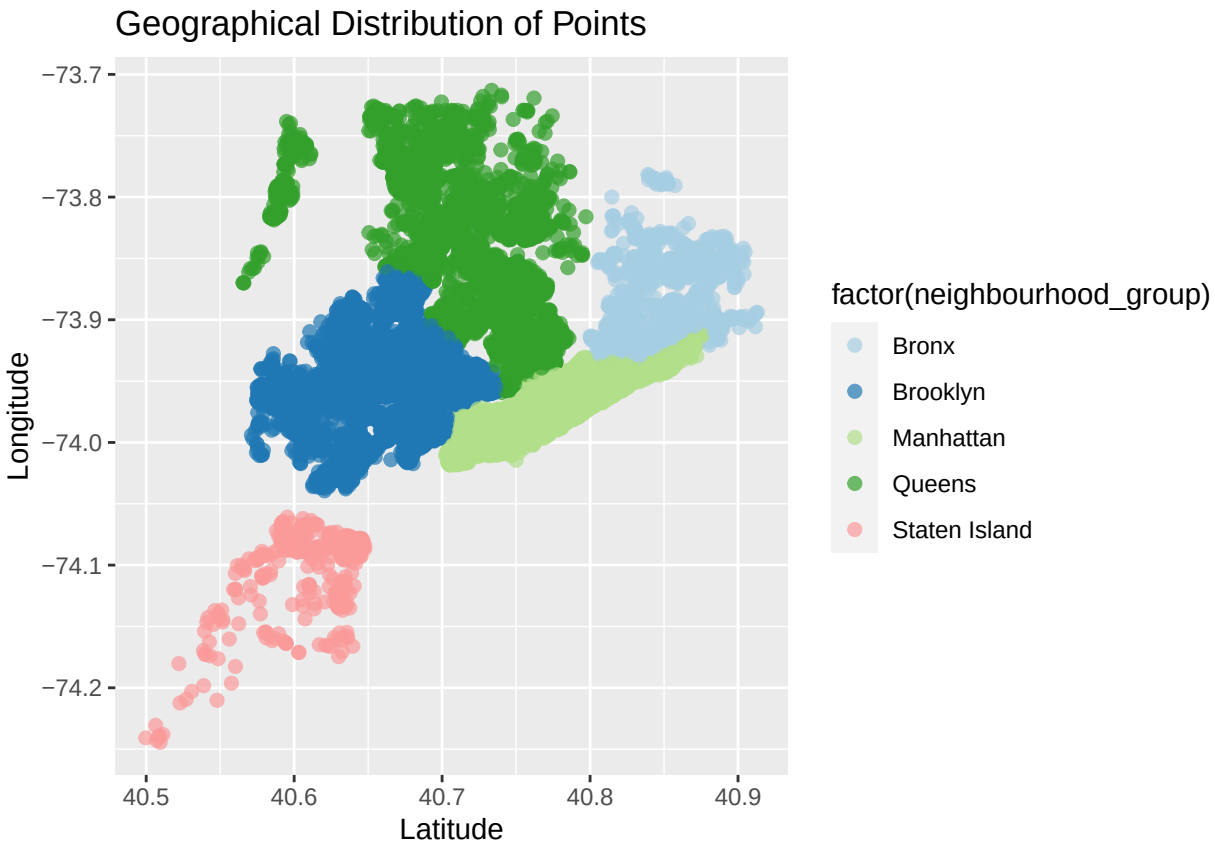
```
cor(NY_data$number_of_reviews,NY_data$price)
```

```
## [1] -0.04795423
```

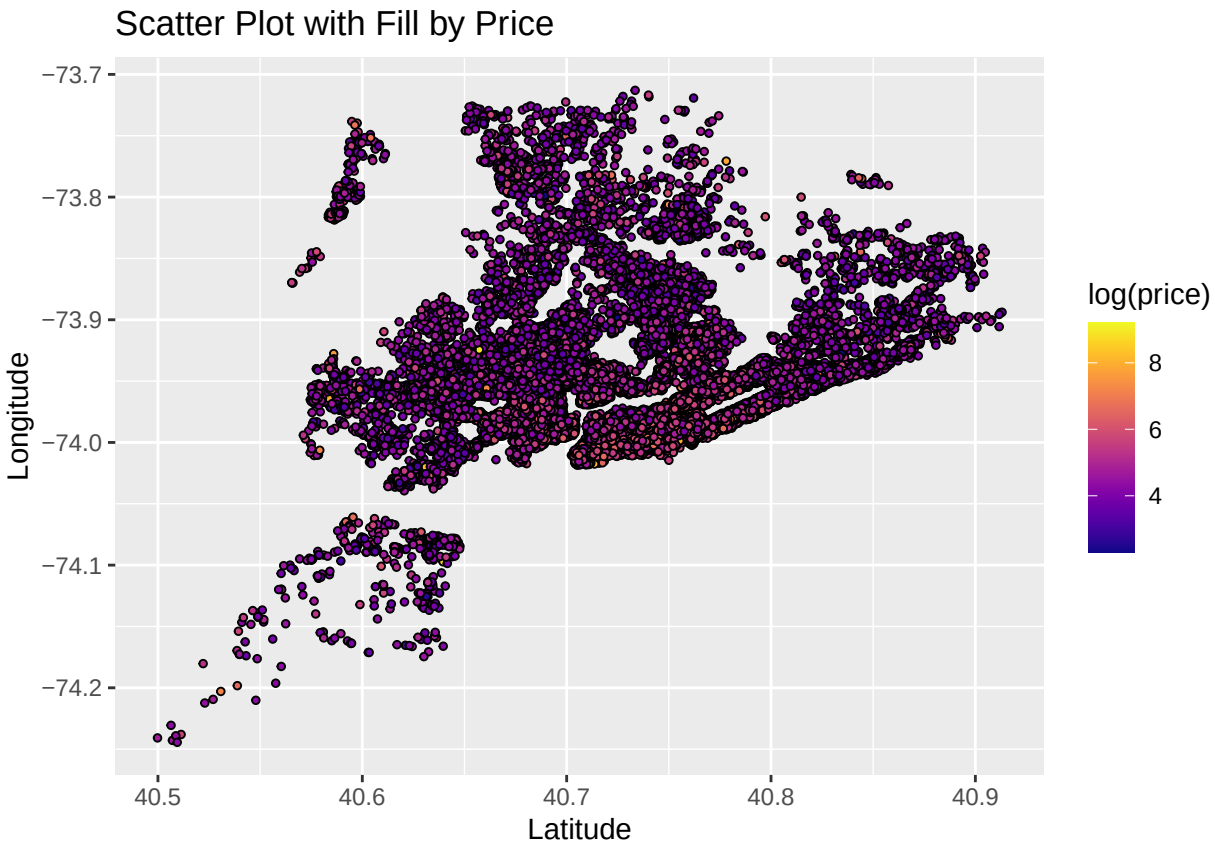
5 Does geographic location influence the price of the apartment?

6 Areas more costly than others

```
ggplot(NY_data, aes(x = latitude, y = longitude, color = factor(neighbourhood_group))) +
  geom_point(alpha = 0.7, size = 2) +
  scale_color_brewer(palette = "Paired") +
  labs(title = "Geographical Distribution of Points",
       x = "Latitude",
       y = "Longitude")
```




```
ggplot(NY_data, aes(x = latitude, y = longitude, fill = log(price))) +
  geom_point(shape = 21, size = 1, color = "black") +
  scale_fill_viridis(option = "plasma", direction = 1) +
  labs(title = "Scatter Plot with Fill by Price",
       x = "Latitude",
       y = "Longitude")
```



- Why a data-driven, computational approach may be useful to answer the questions?
- Describe the dataset that you choose.

Computational Methods [30 points]

- For the choosen dataset, what are the necessary data wrangling steps to make the data ready for subsequent analyses?
- What exploratory analyses and modeling techniques can be used to answer the research questions?
- What metrics will be used to evaluate the quality of the data analysis?

Data Analysis and Results [40 points]

- Perform data analysis, document the analysis procedure, and evaluate the outcomes.
- Present the data analysis results.
- Interpret the results in a way to address the research questions.

Conclusion [15 points]

- Does the analysis answer the research questions?
- Discuss the scope and generalizability of the analysis.
- Discuss potential limitations and possibilities for improvement.