Data606 project proposal

angel

2024-04-04

Data Preparation

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scales)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## 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 error
```

```
# load data
df <- read.csv("https://raw.githubusercontent.com/Angelogallardo05/Data606-proposal/main/NY-House-Datas
```

```
df <- na.omit(df)
```

Research question

Can the amount of square footage, rooms, bathrooms, longitude, and latitude predict a home price in NYC?

Cases

There is data on about 5K homes for sale in NYC

Data collection

Kaggle

Type of study

the effect of sq footage, beds, bathrooms, in home prices.

Data Source

https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market

Dependent Variable

home market price, quantitative

Independent Variable(s)

square footage, bathrooms, rooms. borough

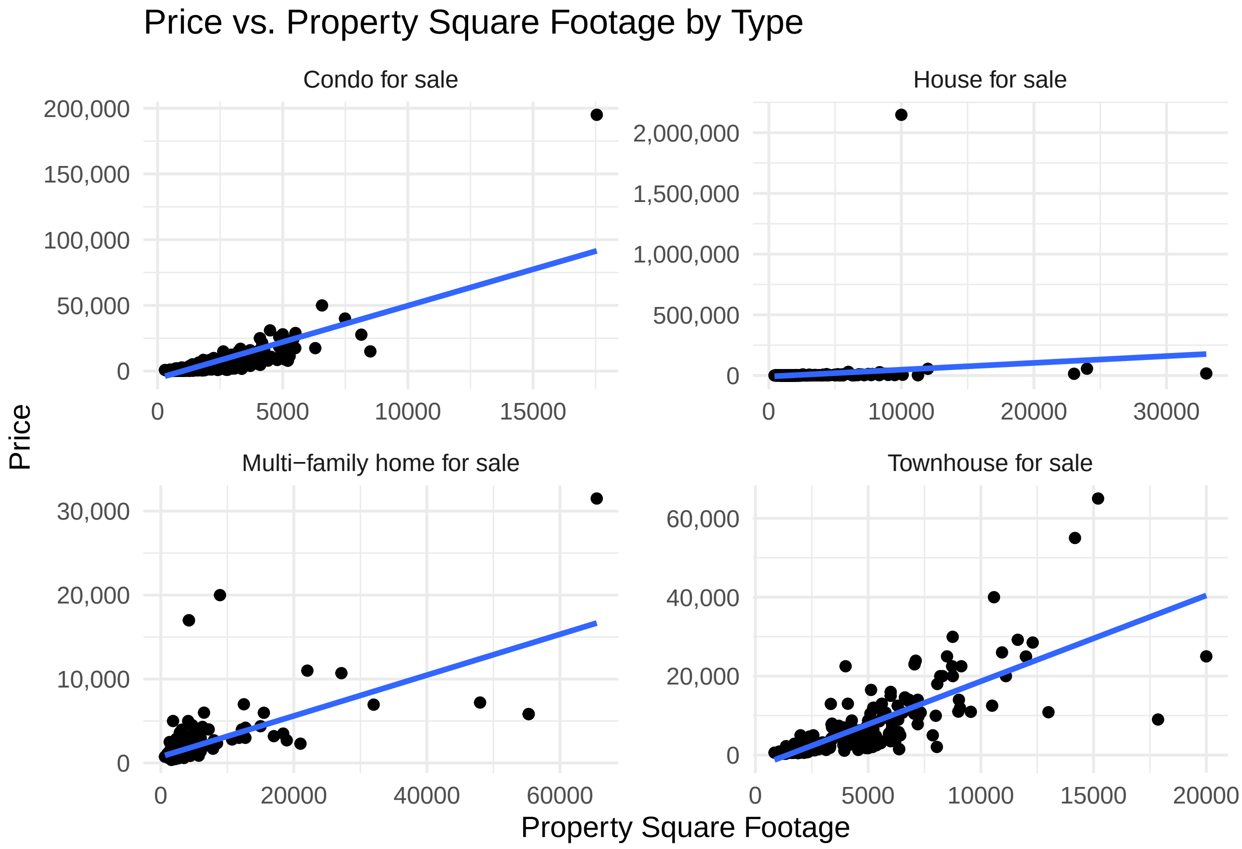
Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```r
library(dplyr)
df_cleaned <- df %>%
  na.omit() %>%
  filter(PROPERTYSQFT != 2184.207862) %>%
   filter(!is.na(PRICE), !grepl("[^0-9.]", PRICE)) %>%
  mutate(PRICE = as.numeric(PRICE))
```

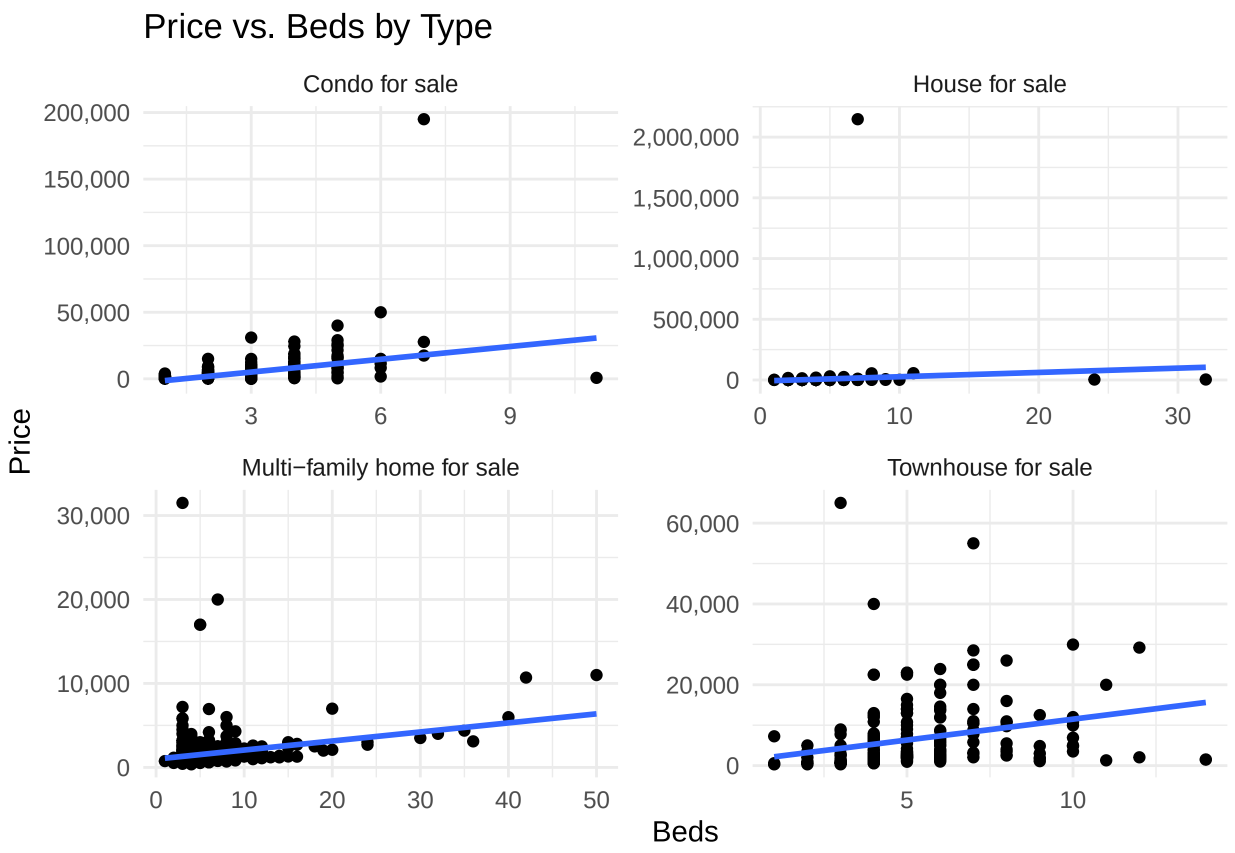
```r
df_grouped <- df_cleaned %>%
  filter(TYPE %in% c("House for sale", "Condo for sale", "Townhouse for sale", "Multi-family home for sa
  select(TYPE, PRICE, BEDS, BATH, PROPERTYSQFT, LONGITUDE, LATITUDE) %>%
  na.omit()


ggplot(df_grouped, aes(x = PROPERTYSQFT, y = PRICE)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Property Square Footage", y = "Price", title = "Price vs. Property Square Footage by Type")
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3)) +
  facet_wrap(~ TYPE, scales = "free") +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



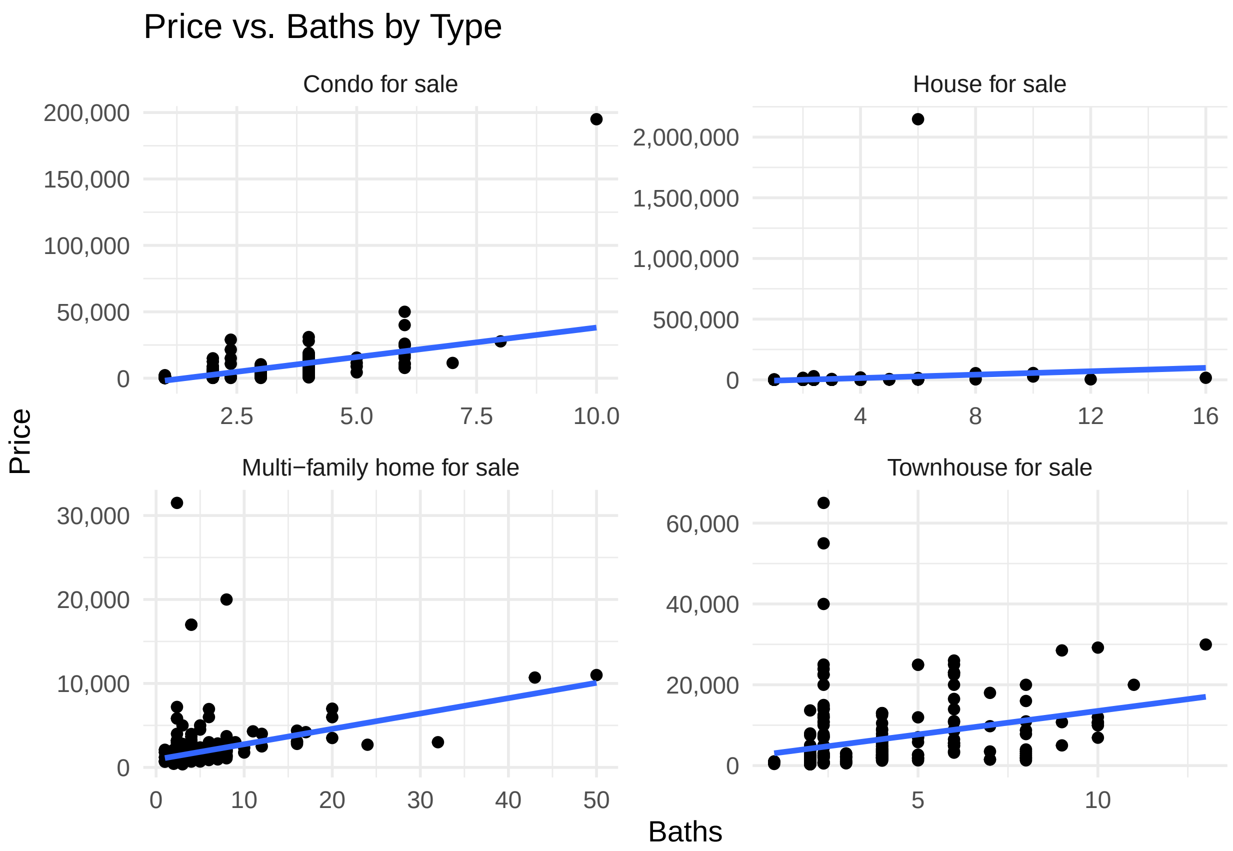
```
ggplot(df_grouped, aes(x = BEDS, y = PRICE)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Beds", y = "Price", title = "Price vs. Beds by Type") +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3)) +
  facet_wrap(~ TYPE, scales = "free") +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



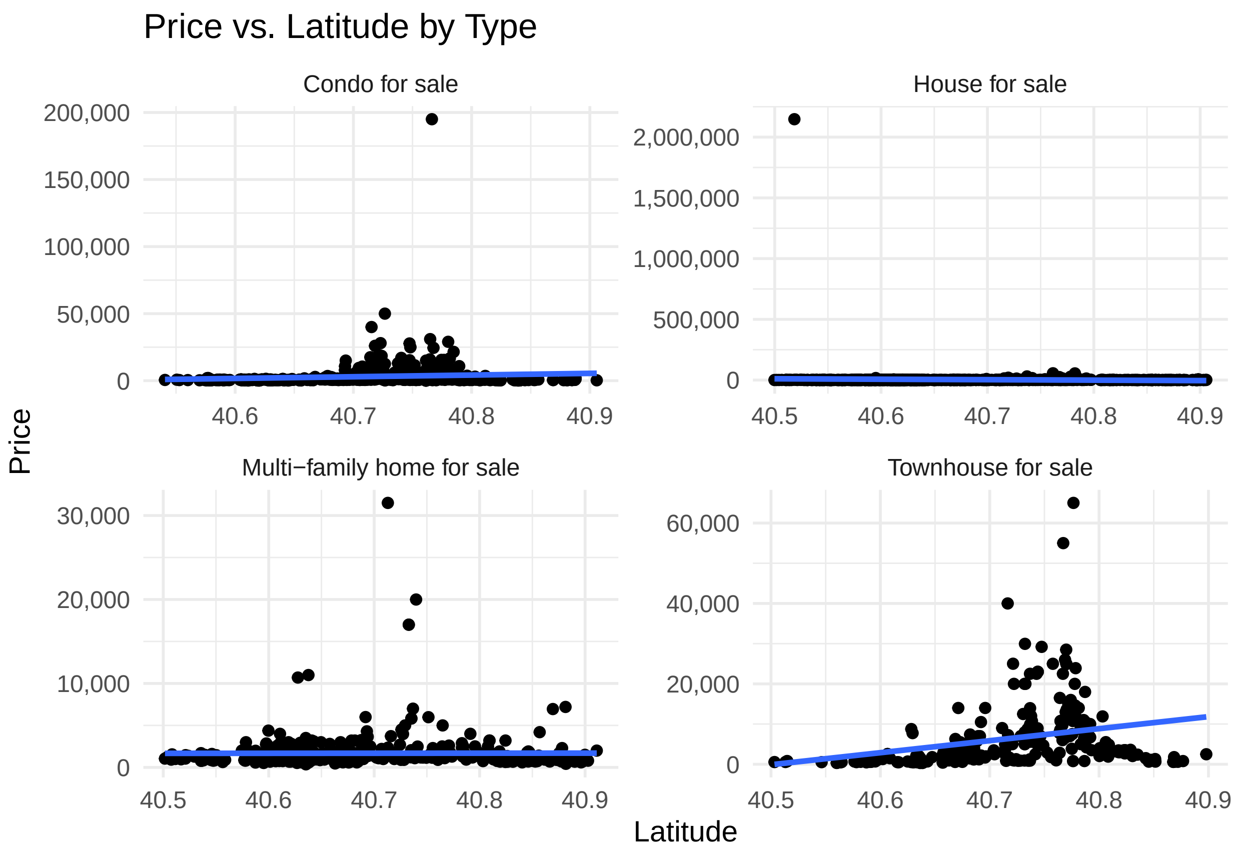
```
ggplot(df_grouped, aes(x = BATH, y = PRICE)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Baths", y = "Price", title = "Price vs. Baths by Type") +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3)) +
  facet_wrap(~ TYPE, scales = "free") +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



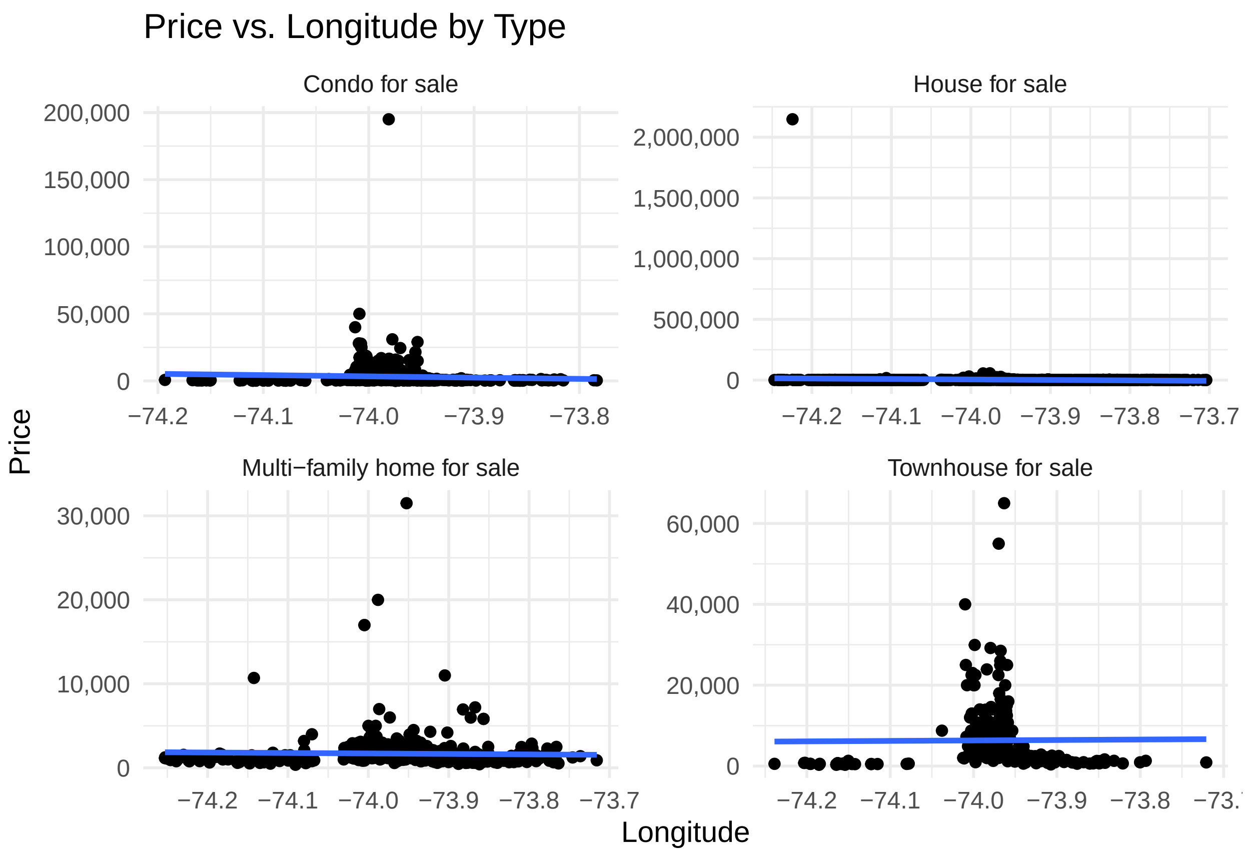
```
ggplot(df_grouped, aes(x = LATITUDE, y = PRICE)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Latitude", y = "Price", title = "Price vs. Latitude by Type") +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3)) +
  facet_wrap(~ TYPE, scales = "free") +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

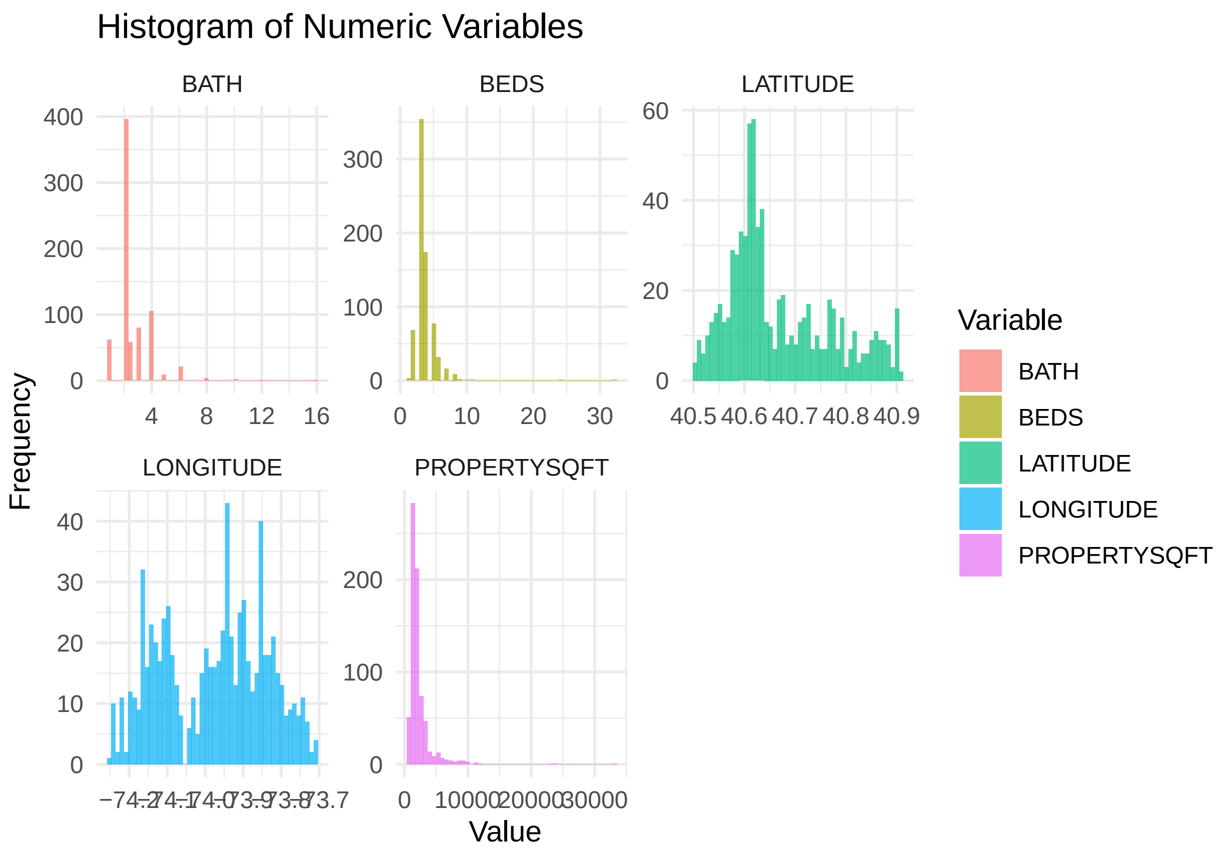


```
ggplot(df_grouped, aes(x = LONGITUDE, y = PRICE)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Longitude", y = "Price", title = "Price vs. Longitude by Type") +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3)) +
  facet_wrap(~ TYPE, scales = "free") +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
selected_cols <- df_cleaned %>%
  filter(TYPE == 'House for sale') %>%
  select(BEDS, BATH, PROPERTYSQFT, LATITUDE, LONGITUDE) %>%
  na.omit()


selected_cols_long <- selected_cols %>%
  pivot_longer(cols = c(BEDS, BATH, PROPERTYSQFT, LATITUDE, LONGITUDE),
               names_to = "Variable", values_to = "Value")


ggplot(selected_cols_long, aes(x = Value, fill = Variable)) +
  geom_histogram(bins = 50, alpha = 0.7, position = "identity") +
  facet_wrap(~ Variable, scales = "free") +
  labs(title = "Histogram of Numeric Variables",
       x = "Value", y = "Frequency") +
  theme_minimal()
```


Histogram of Numeric Variables
BATH
BEDS
LATITUDE
LONGITUDE
PROPERTYSQFT
Frequency
Value
400
300
200
100
0
4
8
12
16
300
200
100
0
0
10
20
30
60
40
20
0
40.5
40.6
40.7
40.8
40.9
40
30
20
10
0
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
200
100
0
0
10000
20000
30000
Variable
BATH
BEDS
LATITUDE
LONGITUDE
PROPERTYSQFT