testing

Stephanie Cheng

2025-04-08

```
##Getting the datasets from the Excel files
manhattan_data<-readxl::read_xlsx(here::here("data","rollingsales_manhattan.xlsx")) |>janitor::clean_na
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...21`
```

```
bronx_data<-readxl::read_xlsx(here::here("data","rollingsales_bronx.xlsx")) |>janitor::clean_names()
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
```

```
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...21`
```

```
brooklyn_data<-readxl::read_xlsx(here::here("data","rollingsales_brooklyn.xlsx")) |>janitor::clean_names
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...21`
```

```
queens_data<-readxl::read_xlsx(here::here("data","rollingsales_queens.xlsx")) |>janitor::clean_names()
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...21`
```

```
staten_island_data<-readxl::read_xlsx(here::here("data","rollingsales_statenisland.xlsx")) |>janitor::cl
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...21`
```

```
##Column Names are the 4th row
col_names<-make.names(c(manhattan_data[4,]))

##The Information in row 1 to 3 are descriptions of the dataset, and are not needed in the dataframe, t

manhattan_data<-manhattan_data[-c(1:4),]
bronx_data<-bronx_data[-c(1:4),]
brooklyn_data<-brooklyn_data[-c(1:4),]
queens_data<-queens_data[-c(1:4),]
staten_island_data<-staten_island_data[-c(1:4),]

##The column names are the same for all datasets, so we can use the same vector for all of them
colnames(manhattan_data)<-col_names
colnames(bronx_data)<-col_names
colnames(brooklyn_data)<-col_names
colnames(queens_data)<-col_names
colnames(staten_island_data)<-col_names
```

```
clean_data <- function(df) {
  df %>%
    filter(SALE.PRICE != 0) %>%
    filter(!is.na(TOTAL.UNITS)) %>%
    select(-NEIGHBORHOOD, -BLOCK, -LOT, -EASEMENT, -ADDRESS, -APARTMENT.NUMBER, -SALE.DATE) %>%
    mutate(
      BOROUGH= as.numeric(as.factor(BOROUGH)),
      BUILDING.CLASS.CATEGORY= as.numeric(as.factor(BUILDING.CLASS.CATEGORY)),
      TAX.CLASS.AT.PRESENT= as.numeric(as.factor(TAX.CLASS.AT.PRESENT)),
      BUILDING.CLASS.AT.PRESENT= as.numeric(as.factor(BUILDING.CLASS.AT.PRESENT)),
      ZIP.CODE= as.numeric(as.character(ZIP.CODE)),
      RESIDENTIAL.UNITS= as.numeric(RESIDENTIAL.UNITS),
```

```
      COMMERCIAL.UNITS= as.numeric(COMMERCIAL.UNITS),
      TOTAL.UNITS= as.numeric(TOTAL.UNITS),
      LAND.SQUARE.FEET= as.numeric(LAND.SQUARE.FEET),
      GROSS.SQUARE.FEET= as.numeric(GROSS.SQUARE.FEET),
      YEAR.BUILT= as.numeric(YEAR.BUILT),
      TAX.CLASS.AT.TIME.OF.SALE= as.numeric(as.factor(TAX.CLASS.AT.TIME.OF.SALE)),
      BUILDING.CLASS.AT.TIME.OF.SALE= as.numeric(as.factor(BUILDING.CLASS.AT.TIME.OF.SALE)),
      SALE.PRICE = as.numeric(SALE.PRICE)
    )
}
manhattan_data <- clean_data(manhattan_data)
bronx_data <- clean_data(bronx_data)
brooklyn_data <- clean_data(brooklyn_data)
queens_data <- clean_data(queens_data)
staten_island_data <- clean_data(staten_island_data)
```