Data Analysis

Wisdom

2023-04-09

```
my_dataset <- read.csv('new.csv', fileEncoding = 'latin1')
```

```
library(vtable)
```

```
## Loading required package: kableExtra
```

```
library(magrittr)
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following object is masked from ’package:kableExtra’:
##
##     group_rows

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

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

```
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(ggrepel)
```

```
## Loading required package: ggplot2
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.2.1     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()    masks magrittr::extract()
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
## x purrr::set_names()  masks magrittr::set_names()
```

```r
library(ggtext)
my_dataset <- my_dataset %>%
  select(-c(1:5,16))
my_dataset <- na.omit(my_dataset)

#my_dataset$tradeTime <- as.numeric(my_dataset$tradeTime)
my_dataset$followers <- as.numeric(my_dataset$followers)
my_dataset$price <- as.numeric(my_dataset$price)
my_dataset$livingRoom <- as.numeric(my_dataset$livingRoom)
my_dataset$drawingRoom <- as.numeric(my_dataset$drawingRoom)
my_dataset$kitchen <- as.numeric(my_dataset$kitchen)
my_dataset$bathRoom <- as.numeric(my_dataset$bathRoom)
my_dataset$constructionTime <- as.numeric(my_dataset$constructionTime)
```

```
## Warning: NAs introduced by coercion
```

```r
my_dataset$renovationCondition <- as.numeric(my_dataset$renovationCondition)
my_dataset$buildingStructure <- as.numeric(my_dataset$buildingStructure)
my_dataset$district <- as.numeric(my_dataset$district)
```

#Renaming variables

```r
colnames(my_dataset)
```

```
##  [1] "tradeTime"           "DOM"                 "followers"
##  [4] "totalPrice"          "price"               "square"
##  [7] "livingRoom"          "drawingRoom"         "kitchen"
## [10] "bathRoom"            "buildingType"        "constructionTime"
## [13] "renovationCondition" "buildingStructure"   "ladderRatio"
## [16] "elevator"            "fiveYearsProperty"   "subway"
## [19] "district"            "communityAverage"
```

```r
my_dataset <-my_dataset %>%
  rename(
    'Trade Time' = tradeTime,
    'Total Price' = totalPrice,
    'Living Room' = livingRoom,
    'Drawing Room' = drawingRoom,
    'Bath Room' = bathRoom,
    'Building Type' = buildingType,
    'Construction Time' = constructionTime,
    'Renovation Condition' = renovationCondition,
    'Building Structure' = buildingStructure,
    'Ladder Ratio' = ladderRatio,
    'Five Years Property' = fiveYearsProperty,
    'Community Average' = communityAverage,
    'Kitchen' = kitchen,
    'Price' = price,
    'Followers' = followers,
    'Square' = square,
    'Elevator' =elevator,
    'Subway' = subway,
    'District' = district
  )
```

```
top_features <- my_dataset[,2:11] # for correlation
str(top_features)
```

```
## ’data.frame’:    159376 obs. of  10 variables:
##  $ DOM          : num  1464 903 1271 965 927 ...
##  $ Followers    : num  106 126 48 138 286 57 167 138 218 134 ...
##  $ Total Price  : num  415 575 1030 298 392 ...
##  $ Price        : num  31680 43436 52021 22202 48396 ...
##  $ Square       : num  131 132 198 134 81 ...
##  $ Living Room  : num  2 2 3 3 2 1 2 3 1 1 ...
##  $ Drawing Room : num  1 2 2 1 1 0 1 2 0 0 ...
##  $ Kitchen      : num  1 1 1 1 1 1 1 1 1 0 ...
##  $ Bath Room    : num  1 2 3 1 1 1 1 2 1 0 ...
##  $ Building Type: num  1 1 4 1 4 4 4 1 3 1 ...
```

```
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: ’reshape2’
```

```
## The following object is masked from ’package:tidyr’:
##
##     smiths
```

```
dt <- melt(cor(top_features, use="p"))
dt$value <- trunc(dt$value*10^2)/10^2
heat1 <- ggplot(data = dt, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  geom_text(aes(Var2, Var1, label = value), size = 5) +
  scale_fill_gradient2(low = "red", high = "green",
                       limit = c(-1,1), name="Correlation") +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.background = element_blank())

png(file = "Correlation_heatmap.png")
heat1 + theme(axis.title = element_blank())
```

```
#summary statistics
sum_data <- top_features[,3:10]
```

```
library(tidyr)
sum_dt <- sum_data %>%
  pivot_longer(names_to = 'House rooms', values_to = 'Total rooms',cols = -c(`Total Price`,Square,Price
```

```
# Plot the chart.
set.seed(1234)
data <- sum_dt %>% mutate(`Building Type` = case_when(
  `Building Type` == 1 ~ "Tower",
  `Building Type` == 2 ~ "Bungalow",
  `Building Type` == 3 ~ "Combination",
  `Building Type` == 4 ~ "Plate",
  TRUE ~ as.character(`Building Type`)
  ))

top_feat <- data %>% sample_n(1000)
top_feat <- data

graph2 <-top_feat %>%
  ggplot(aes(x= `Building Type`, y = `Total Price`, fill = 'none'))+
  geom_col()+
  ggtitle('Building Type by Price')


options(scipen = 999)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
png(file = " Prices per square and total rooms.png")
graph2 + theme(legend.position = 'none') +
  labs(x = 'Building Types', y = 'Total Prices')+
  scale_y_continuous(labels = function(`Total Price`) paste0("$", format(`Total Price`, big.mark = ","))
    theme(plot.title = element_markdown(face = "bold", size = rel(1.6)),
        plot.subtitle = element_markdown(size = rel(1.3)),
        plot.margin = unit(c(0.5, 1, 0.5, 0.5), units = "lines"))

ggsave("Building_Type.pdf", graph2,
       width = 8, height = 5, units = "in", device = cairo_pdf)
```

#House price based on number of rooms

```
top_feat$`Total rooms` <- as.factor(top_feat$`Total rooms`)
graph3 <- top_feat %>%
  ggplot(aes(x= Price, color = `Total rooms`))+geom_density()+
  labs(title = 'Total House price based on number of rooms', y = 'price rise')


graph_3 <- graph3 + labs(x ='Range of price per square(metre)', y = 'House price fluctuations' ) + theme

options(scipen = 999)
library(scales)
png(file = "Total Rooms influence on Price.png")
graph_3
```

```
str(top_feat)
```

```
## tibble [637,504 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Total Price  : num [1:637504] 415 415 415 415 575 575 575 575 1030 1030 ...
##  $ Price        : num [1:637504] 31680 31680 31680 31680 43436 ...
##  $ Square       : num [1:637504] 131 131 131 131 132 ...
##  $ Building Type: chr [1:637504] "Tower" "Tower" "Tower" "Tower" ...
##  $ House rooms  : chr [1:637504] "Living Room" "Drawing Room" "Kitchen" "Bath Room" ...
##  $ Total rooms  : Factor w/ 8 levels "0","1","2","3",..: 3 2 2 2 3 3 2 3 4 3 ...
```

```
top_feat$Building_type <- with(top_feat, ifelse(top_feat$`Building Type` == 1,'tower',
                                                ifelse(top_feat$`Building Type` == 2,'bungalow',
                                                       ifelse(top_feat$`Building Type` == 3, 'plate & to

top_feat
```

```
## # A tibble: 637,504 x 7
##    ‘Total Price‘ Price Square ‘Building Type‘ ‘House rooms‘ ‘Total rooms‘
##            <dbl> <dbl>  <dbl> <chr>           <chr>         <fct>
##  1           415 31680   131  Tower           Living Room   2
##  2           415 31680   131  Tower           Drawing Room  1
##  3           415 31680   131  Tower           Kitchen       1
##  4           415 31680   131  Tower           Bath Room     1
##  5           575 43436   132. Tower           Living Room   2
##  6           575 43436   132. Tower           Drawing Room  2
##  7           575 43436   132. Tower           Kitchen       1
##  8           575 43436   132. Tower           Bath Room     2
##  9          1030 52021   198  Plate           Living Room   3
## 10          1030 52021   198  Plate           Drawing Room  2
## # i 637,494 more rows
## # i 1 more variable: Building_type <chr>
```

```
graph4 <-top_feat %>%
  ggplot(aes(x= `Building Type`, y = `Total Price`))+
  geom_col()+
  ggtitle('Building type based on price')

graph_4 <- graph4 + theme(legend.position = 'none')

options(scipen = 999)
library(scales)
png(file = "Building type based on price.png")
graph_4
```

```
#linear regression
LN_regression <- lm(top_features$Price~top_features$Square)

LN_regression2 <- lm(top_features$Price~top_features$`Building Type`)

LN_reg3 <- lm(top_features$Price~top_features$Square+top_features$`Building Type`)
LN_reg4 <- lm(top_features$Price~top_features$Square+top_features$Kitchen+top_features$`Building Type`+
stargazer(LN_regression,LN_regression2,LN_reg3,LN_reg4, type= 'text', out = 'Regression table')
```

```
##
## ====================================================================================================
##                                                                       Dependent variable:
##                      --------------------------------------------------------------------------------
##                                                                             Price
##                                   (1)                          (2)                          (3)
## ----------------------------------------------------------------------------------------------------
## Square                         -118.695***                                               -119.239***
##                                  (1.623)                                                   (1.621)
##
## Kitchen
##
##
## ‘Building Type‘                                           -895.429***                    -952.624***
##                                                             (47.742)                     (46.958)
##
## ‘Drawing Room‘
##
##
## ‘Living Room‘
##
##
## ‘Bath Room‘
##
##
## DOM
##
##
```

```
## Constant                    61,251.780***                54,154.920***                64,176.200***
##                               (146.594)                    (156.410)                    (205.463)
##
## ----------------------------------------------------------------------------------------------------
## Observations                   159,376                      159,376                      159,376
## R2                              0.032                        0.002                        0.035
## Adjusted R2                     0.032                        0.002                        0.035
## Residual Std. Error   23,715.900 (df = 159374)     24,084.170 (df = 159374)     23,685.410 (df = 159
## F Statistic         5,350.983*** (df = 1; 159374) 351.773*** (df = 1; 159374) 2,888.160*** (df = 2;
## ====================================================================================================
## Note:
```

```
summary_table <- sumtable(sum_data,
        summ = c('mean(x)',
                 'sd(x)','min(x)',
                 'max(x)'),
        title = 'Summary Statistics House Prices',
        out = 'return')

summary_table
```

```
##        Variable  Mean    Sd Min    Max
## 1   Total Price   409   254 0.1   4900
## 2         Price 51448 24111   1 156250
## 3        Square    83    37 7.4    640
## 4   Living Room     2  0.77   0      7
## 5  Drawing Room   1.1  0.51   0      5
## 6       Kitchen  0.99  0.12   0      3
## 7     Bath Room   1.2  0.43   0      6
## 8 Building Type     3   1.3   1      4
```