Assignment 1

Chen Xinyu A0198779W

2020/2/17

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --
```

```
## √ ggplot2 3.2.1     √ purrr   0.3.3
## √ tibble  2.1.3     √ dplyr   0.8.3
## √ tidyr   1.0.0     √ stringr 1.4.0
## √ readr   1.3.1     √ forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflic
ts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dbplyr)
```

```
##
## Attaching package: 'dbplyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     ident, sql
```

```
library(ggplot2)
library(ggridges)
```

Task 1 HDB RESALE FLAT PRICES

1.1 Preparation

1.1.1 Data import

```
data1<-read.csv("resale-flat-prices-based-on-approval-date-1990-1999.csv")
data2<-read.csv("resale-flat-prices-based-on-approval-date-2000-feb-2012.cs
v")
data3<-read.csv("resale-flat-prices-based-on-registration-date-from-mar-2012
-to-dec-2014.csv")
data4<-read.csv("resale-flat-prices-based-on-registration-date-from-jan-2015
-to-dec-2016.csv")
data5<-read.csv("resale-flat-prices-based-on-registration-date-from-jan-2017
-onwards.csv")
```

1.1.2 Data wrangling

The coloum ‘remaning_lease’ is not existing in data1 to data 3, so i need to remove this coloum in data4 and data5, or data cannot be mergerd among datasets with different columns.

Then I calculated the resale prices per squaremeter in a new coloum named **prices_per_sqm** .

```
ndata4<-data4%>%
  select(-remaining_lease)
ndata5<-data5%>%
  select(-remaining_lease)
data<-rbind(data1,data2,data3,ndata4,ndata5)
data<-data%>%
  mutate(prices_per_sqm=resale_price/floor_area_sqm)
```

I seperated the date into column **year** and column **month** to make the analysis easier.

```
data<-data%>%
  separate(month,sep="-", into=c("year","month"))
```

1.2 General Analysis

```
summary(data)
```

```
##       year              month                    town
##  Length:810795      Length:810795      TAMPINES   : 72008
##  Class :character   Class :character   YISHUN     : 62041
##  Mode  :character   Mode  :character   BEDOK      : 60478
##                                        JURONG WEST: 59671
##                                        WOODLANDS  : 57772
##                                        ANG MO KIO : 47535
##                                        (Other)    :451290
##       flat_type          block                  street_name
##  4 ROOM   :302771   2      :  4262   YISHUN RING RD    : 15966
##  3 ROOM   :268766   1      :  3732   BEDOK RESERVOIR RD: 13521
##  5 ROOM   :166512   110    :  3145   ANG MO KIO AVE 10 : 12755
##  EXECUTIVE: 61440   101    :  3113   ANG MO KIO AVE 3  : 11235
##  2 ROOM   :  9543   4      :  3054   HOUGANG AVE 8     :  8533
##  1 ROOM   :  1265   113    :  3046   TAMPINES ST 21    :  7647
##  (Other)  :   498   (Other):790443   (Other)           :741138
##    storey_range    floor_area_sqm           flat_model
##  04 TO 06:205966   Min.   : 28.00   Model A       :152566
##  07 TO 09:185919   1st Qu.: 73.00   Improved      :139894
##  01 TO 03:166020   Median : 93.00   New Generation: 96681
##  10 TO 12:157418   Mean   : 95.53   NEW GENERATION: 78898
##  13 TO 15: 50669   3rd Qu.:114.00   IMPROVED      : 73593
##  16 TO 18: 18734   Max.   :307.00   MODEL A       : 70381
##  (Other) : 26069                    (Other)       :198782
##  lease_commence_date  resale_price     prices_per_sqm
##  Min.   :1966        Min.   :   5000   Min.   :  161.3
##  1st Qu.:1980        1st Qu.: 180000   1st Qu.: 2219.0
##  Median :1986        Median : 272000   Median : 2747.3
##  Mean   :1987        Mean   : 290692   Mean   : 2999.2
##  3rd Qu.:1993        3rd Qu.: 380000   3rd Qu.: 3714.3
##  Max.   :2016        Max.   :1205000   Max.   :11808.5
##
```

This dataset has 810795 observations and 12 variables in total.

1.3 Plots and Analysis

(1) The changes of HDB resale prices

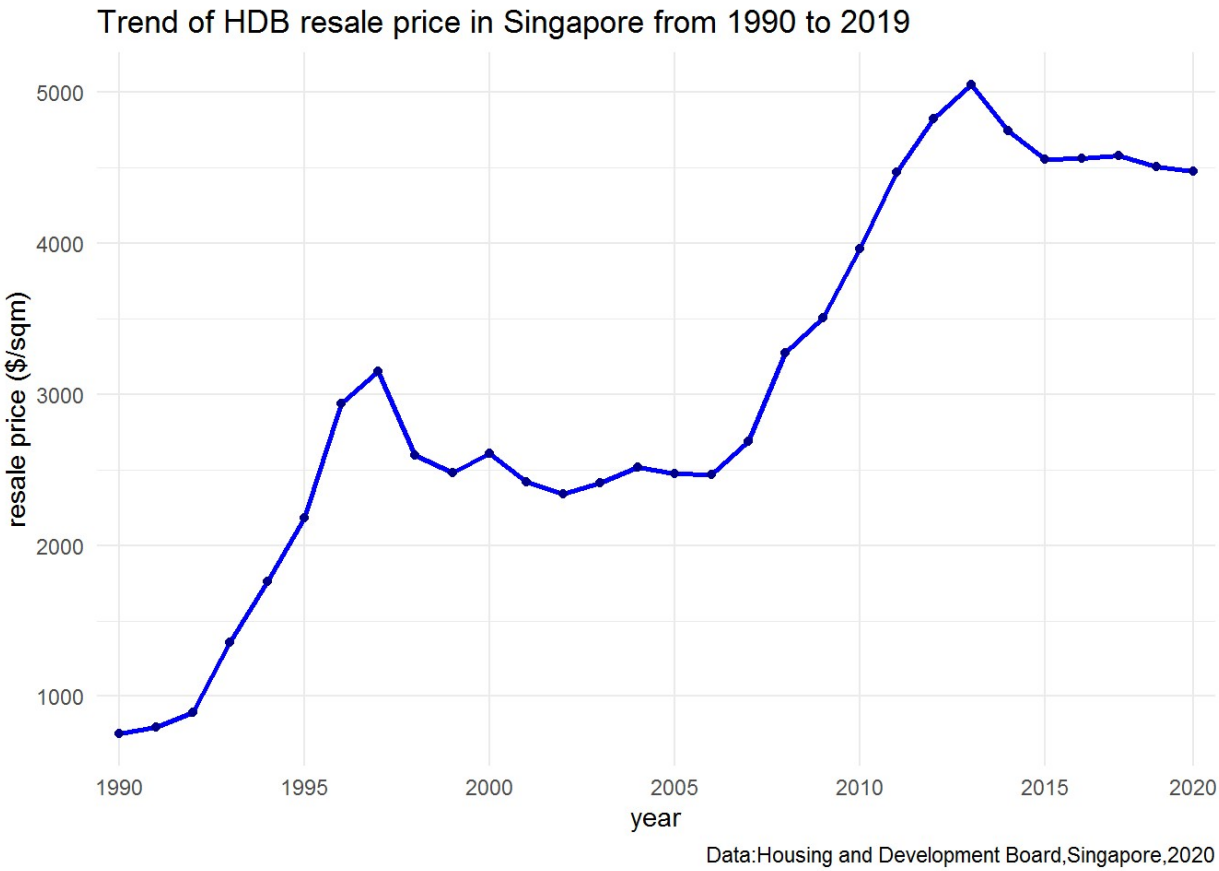
Plot 1

This plot calclulated the mean resale prices ($/sqm) for each year from 1990 to 2019, then shows the trend of it.

```
year_price<-data%>%
  group_by(year)%>%
  summarise(year_price=mean(prices_per_sqm))

ggplot(data=year_price,mapping=aes(x=year,y=year_price))+
  geom_line(color='blue',size=1,group=1,color='light blue')+
 geom_point(color='dark blue')+
  scale_x_discrete(breaks=c(1990,1995,2000,2005,2010,2015,2019),labels=c('19
90','1995','2000','2005','2010','2015','2020'))+
  theme_minimal()+
  labs(y='resale price ($/sqm)',
       x='year',
       title='Trend of HDB resale price in Singapore from 1990 to 2019',
       caption='Data:Housing and Development Board,Singapore,2020')
```

```
## Warning: Duplicated aesthetics after name standardisation: colour
```



Plot 1 shows a general increasing trend of resale prices from 1990 to 2019, with a stable period from 1998 to 2006. In 2013, the price reached its peak.

Plot 2

This plot divides 1990 to 2019 into three decades, and shows the differences of HDB resale prices ($/sqm) in each decades.

```
period_price<-year_price%>%
  mutate(period=case_when(
    year<2000 ~'1990s',
    year>=2000 & year<2010 ~'2000s',
    year>=2000 ~'2010s'
  ))%>%
  group_by(period)%>%
  summarise(period_price=mean(year_price))

ggplot(data=period_price,mapping=aes(x=period,y=period_price))+
  geom_col(fill='light blue',color='blue')+
  theme_minimal()+
  labs(x='Period',
       y='Resale price ($/sqm )',
       title='HDB Resale price in Singapore from 1990s to 2010s',
       caption='Data:Housing and Development Board,Singapore,2020')
```

Plot 2 shows that the mean HDB resale price is always increasing in these three decades. And the change is more drastic between 2010s and 2000s than the one between 2000s and 1990s.

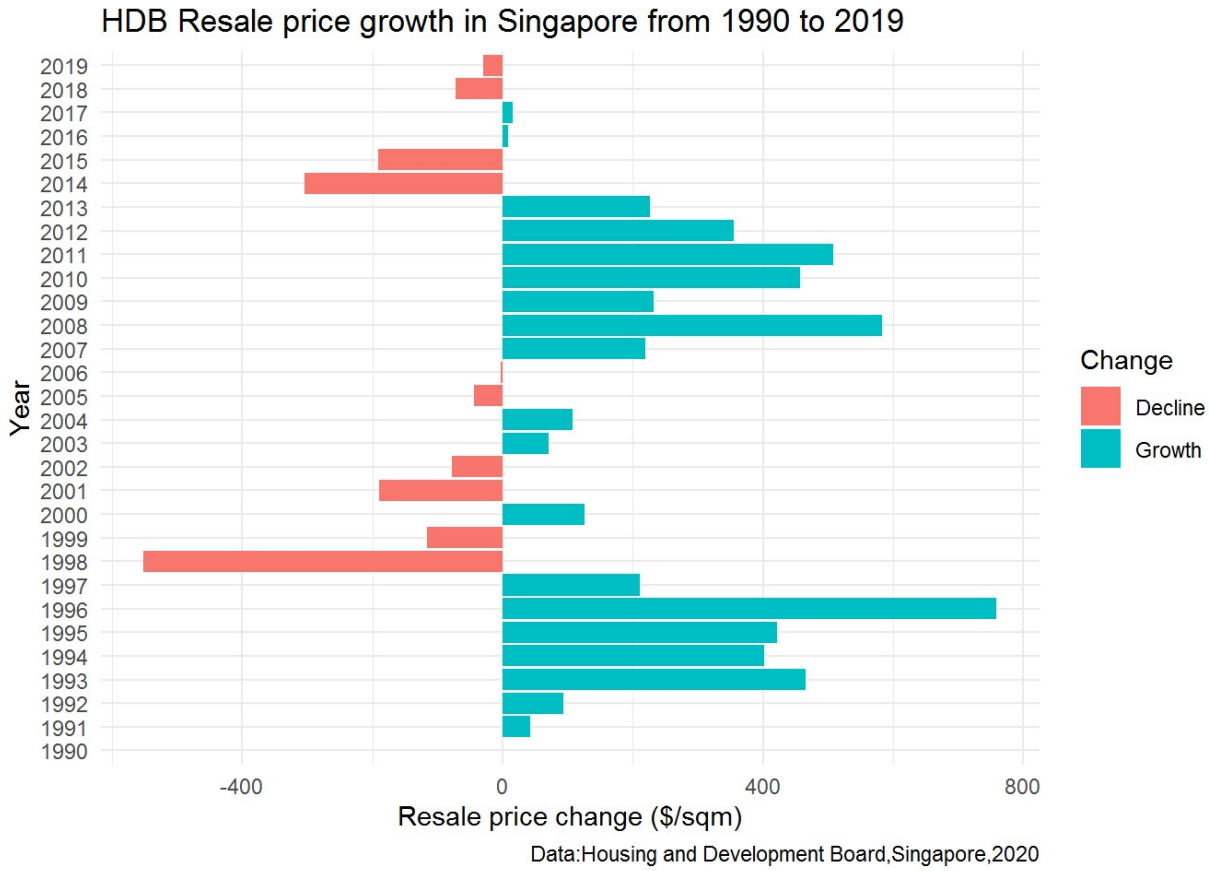
Plot 3

This plot shows the growth of HDB resale price ($/sqm) among years from 1990 to 2019.

```
growth_year_price<-year_price%>%
  mutate(diff=year_price -lag(year_price))%>%
   mutate(Change=case_when(
     diff>0 ~ 'Growth',
     diff<=0 ~'Decline'
   ))

ggplot(data=growth_year_price,mapping=aes(x=year,y=diff,fill=Change))+
  geom_col()+
  coord_flip()+
  theme_minimal()+
  labs(y='Resale price change ($/sqm)',
       x='Year',
       title='HDB Resale price growth in Singapore from 1990 to 2019',
       caption='Data:Housing and Development Board,Singapore,2020')
```

```
## Warning: Removed 1 rows containing missing values (position_stack).
```



Plot 3 shows that there are some fluctuations from 1990 to 2019. 1991-1997,2000,2003-2004,2007-2013 and 2016-2017 are the periods that resale prices growed, conversely,1998-1999,2001-2002,2005-2006,2014-2015 and 2018-2019 are the periods that prices decline. And growth is greater than decline during th 30 years, resulting in a genral upward trend of prices.

(2) The changes of number and types of HDB flats sold from 1990 to 2019

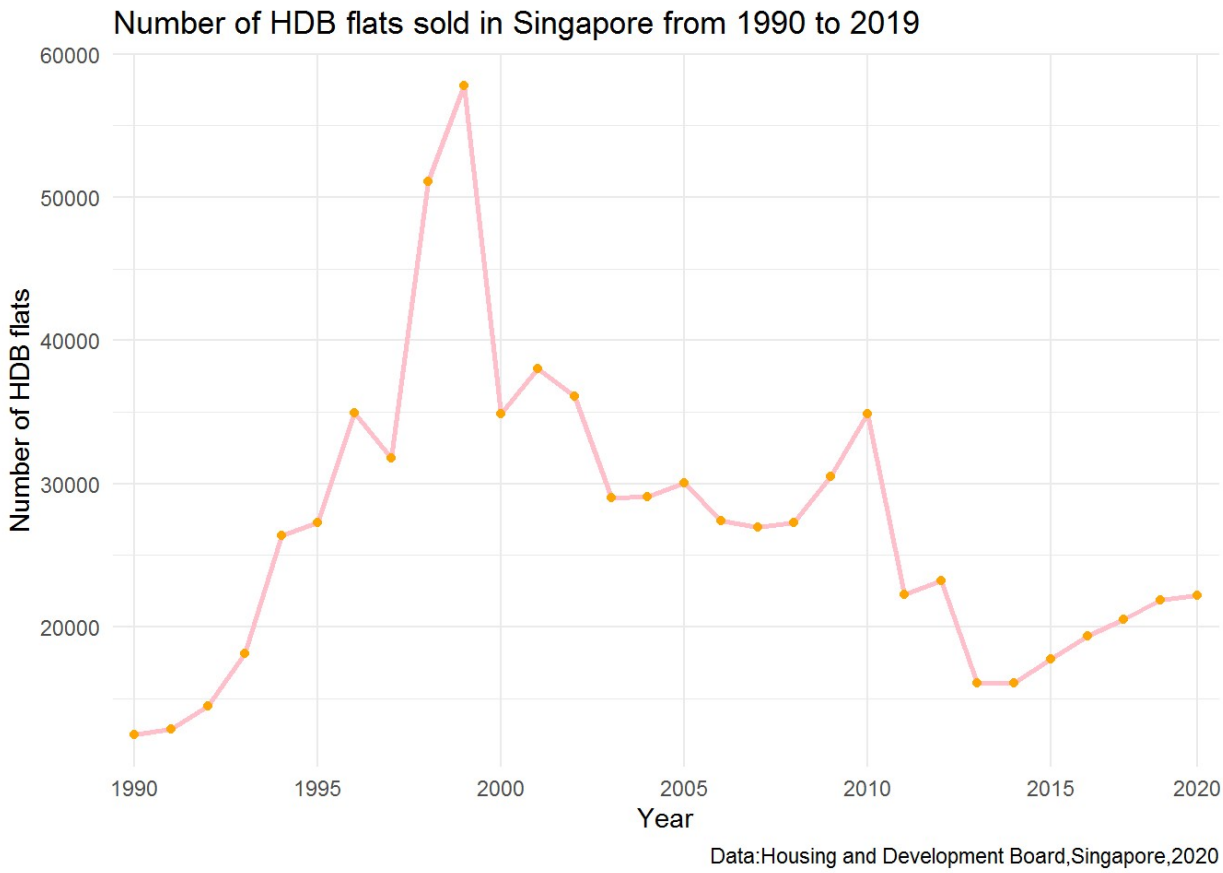
Plot 4

This plot shows the trend of quantity of HDB flats sold in Singapore from 1990 to 2019.

```
(count_year<-data%>%
  group_by(year)%>%
  summarise(count=n()))
```

```
## # A tibble: 30 x 2
##    year  count
##    <chr> <int>
##  1 1990  12505
##  2 1991  12855
##  3 1992  14503
##  4 1993  18116
##  5 1994  26373
##  6 1995  27289
##  7 1996  34919
##  8 1997  31759
##  9 1998  51095
## 10 1999  57786
## # ... with 20 more rows
```

```
ggplot(data=count_year,mapping=aes(x=year,y=count,group=1))+
  geom_line (color='pink',size=1)+
  geom_point(color='orange')+
  scale_x_discrete(breaks=c(1990,1995,2000,2005,2010,2015,2019),labels=c('19
90','1995','2000','2005','2010','2015','2020'))+
  theme_minimal()+
  labs(x='Year',
       y='Number of HDB flats',
      title='Number of HDB flats sold in Singapore from 1990 to 2019',
       caption='Data:Housing and Development Board,Singapore,2020')
```

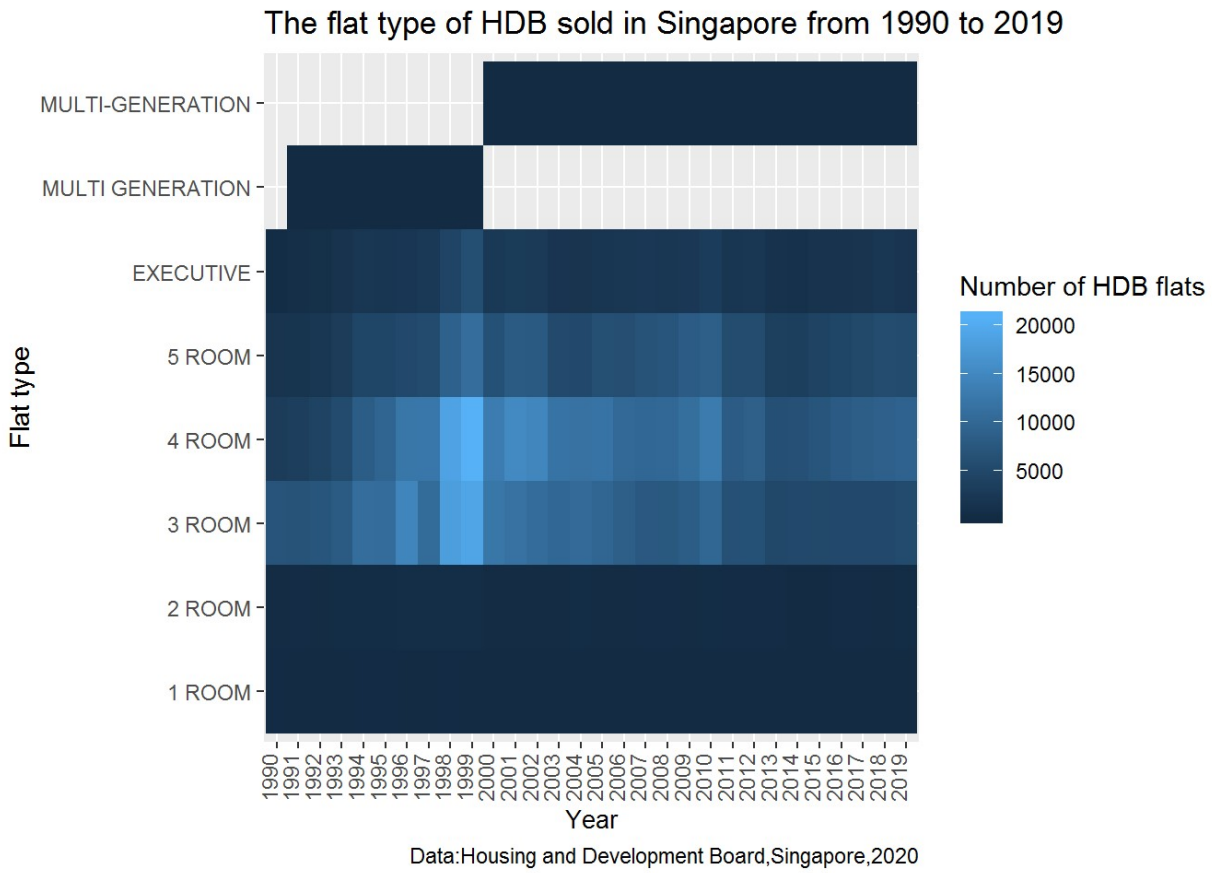


Plot 4 shows a general upward trend from 1990 to 1999, reaching a peak in 1999, then a general downward trend from 2000 to 2019. And these years, the number is increasing gently again.

Plot 5

This plot shows the changes of the flat type of HDB sold in Singapore from 1990 to 2019.

```
ggplot(data=data,mapping=aes(x=year,y=flat_type))+
  geom_bin2d(binwidth=c(1,1))+
  theme(axis.title.x=element_text(size=10),axis.text.x = element_text(angle
= 90,vjust=0))+
  labs(y='Flat type',
       x='Year',
       fill='Number of HDB flats',
       title='The flat type of HDB sold in Singapore from 1990 to 2019',
      caption='Data:Housing and Development Board,Singapore,2020')
```

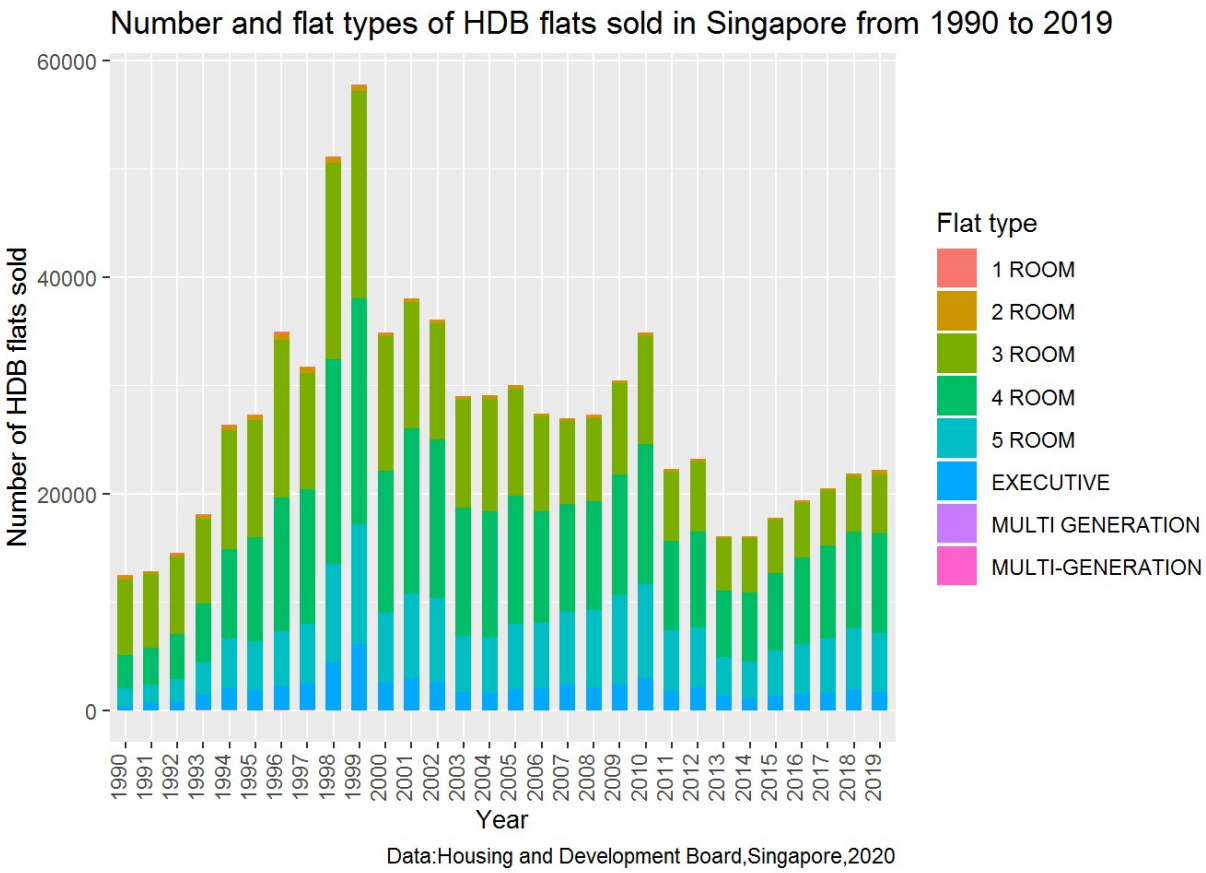


Plot 5 shows that most of the HDB flats are **3-room**, **4-room** and **5-room**. The type of **multi-generation** only existed after year. **3-room** and **4-room** are the most popular types from 1998 to 2000. Nowadays,the distribution among different types of flats sold is more equal than before.

Plot 6

This plot combines the information in plot 4 and plot 5.

```
ggplot(data=data,mapping=aes(x=year,fill=flat_type))+
   geom_bar(width = 0.6,position ="stack")+
  theme(axis.title.x=element_text(size=10),axis.text.x = element_text(angle
= 90,vjust=0))+
 labs(y='Number of HDB flats sold',
       x='Year',
       fill='Flat type',
       title='Number and flat types of HDB flats sold in Singapore from 199
0 to 2019',
      caption='Data:Housing and Development Board,Singapore,2020')
```



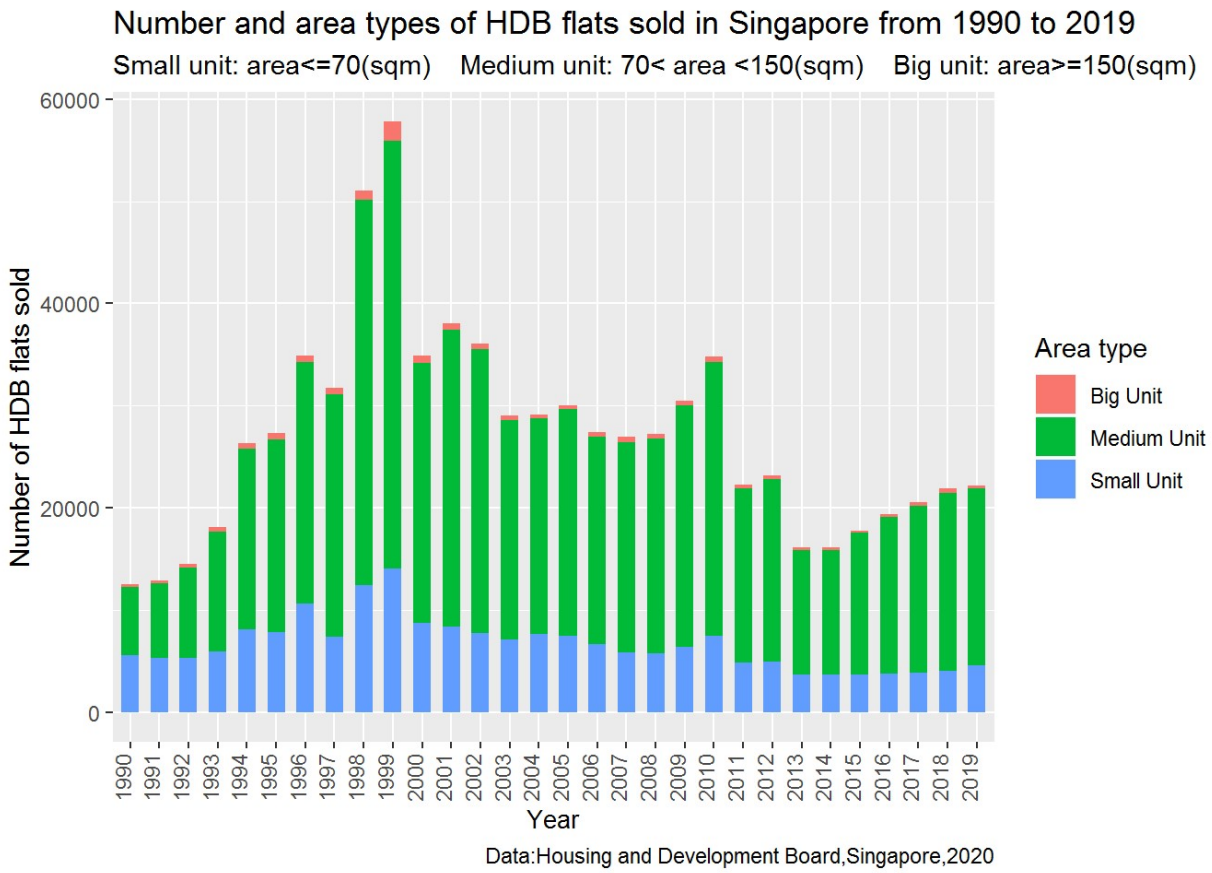
From Plot 6, we can not only see the total number of HDB flats sold in each year, but also can know the quantity relations among different flat types.

Plot 7

This plot shows the total number and numbers of each area types of HDB flats sold in each year.

```
data<-data%>%
  mutate(area_type= case_when(
    floor_area_sqm<=70 ~'Small Unit',
    floor_area_sqm>70 &floor_area_sqm<150~ 'Medium Unit',
    floor_area_sqm>=150~ 'Big Unit'
  ))

ggplot(data=data,mapping=aes(x=year,fill=area_type))+
  geom_bar(width = 0.6,position ="stack")+
  labs(x='Year',
       y='Number of HDB flats sold',
       fill='Area type',
       title='Number and area types of HDB flats sold in Singapore from 199
0 to 2019',
       subtitle='Small unit: area<=70(sqm)    Medium unit: 70< area <150(sq
m)    Big unit: area>=150(sqm)',
       caption='Data:Housing and Development Board,Singapore,2020')+
  theme(axis.title.x=element_text(size=10),axis.text.x = element_text(angle
= 90,vjust=0))
```



In plot 7, we can see the trend of total number of HDB sold in each year as plot 4 shows. Besides, the flats with area between 70 sqm and 150 sqm are sold the most, compared with other two area types.

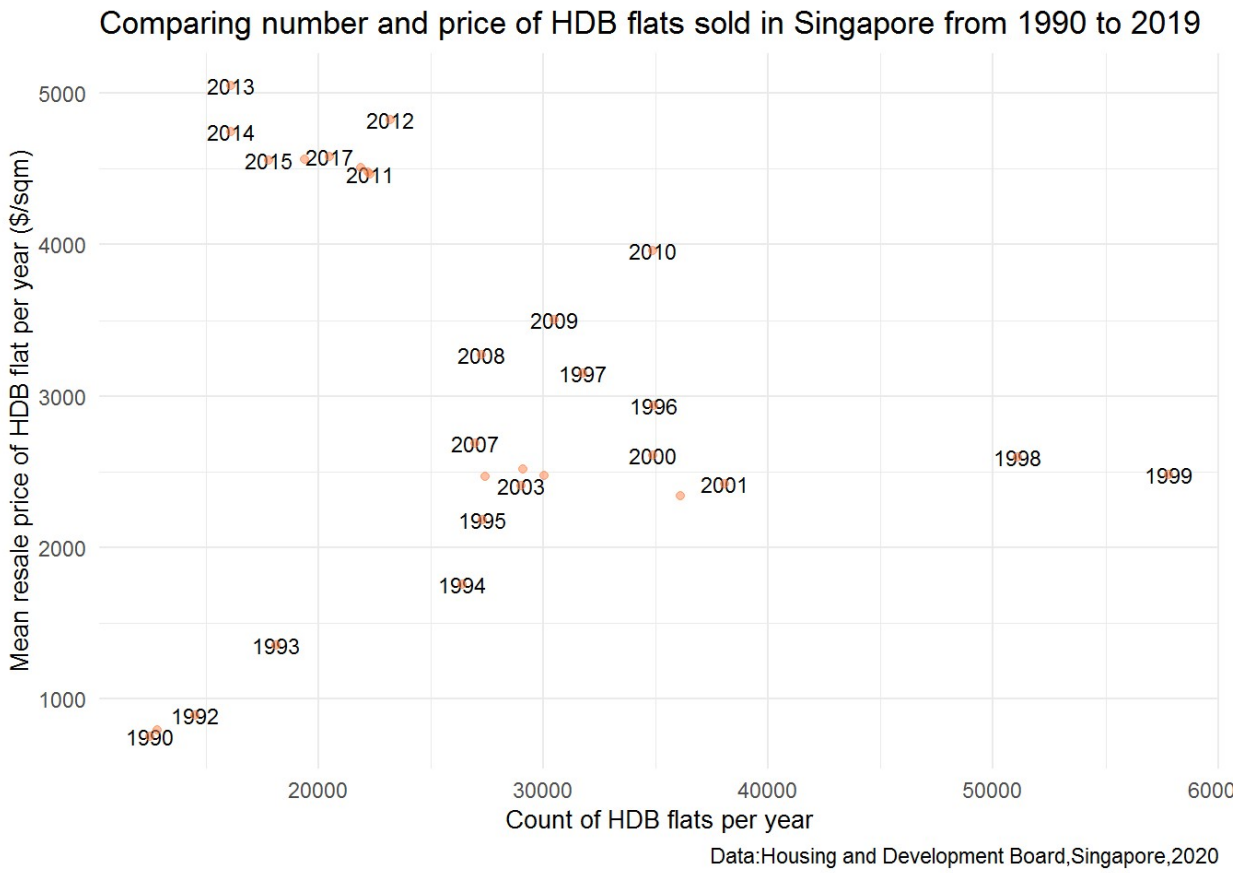
(3) Is there a relationship between number and price of HDB flats sold in Singapore?

Plot 8

The location of the point in this scatter plot shows the number and mean price of HDB flats sold in each year.

```
count_price_year<-cbind(year_price,count_year$count)
names(count_price_year)<-c('year','price','count')
```

```
ggplot(data=count_price_year,mapping=aes(x=count,y=price,label=year))+
  geom_text(check_overlap = TRUE,size=3)+
  geom_point(alpha=0.5,color='sienna1')+
  theme_minimal()+
  labs(x='Count of HDB flats per year',
       y='Mean resale price of HDB flat per year ($/sqm)',
      title='Comparing number and price of HDB flats sold in Singapore from
1990 to 2019',
       caption='Data:Housing and Development Board,Singapore,2020')+
  theme(axis.title=element_text(size=10))
```



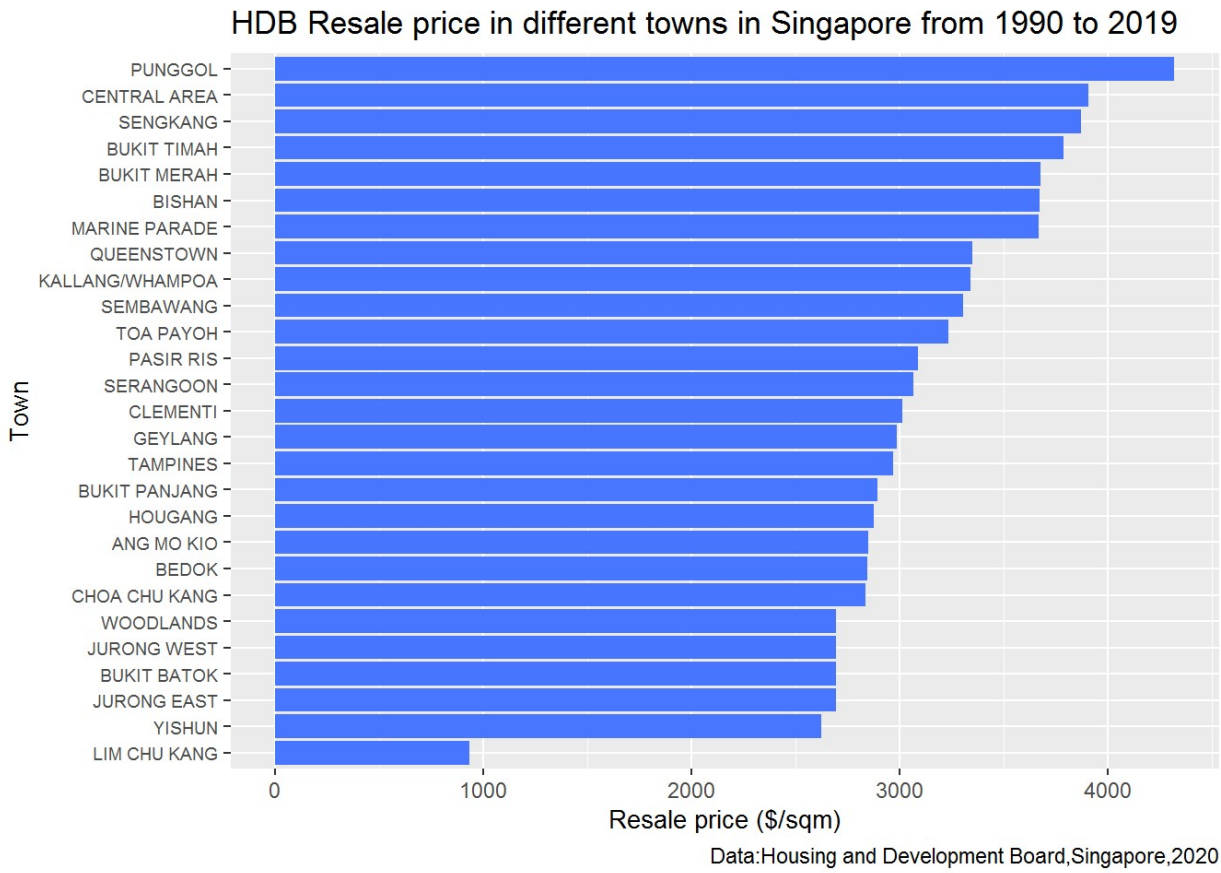
In plot 8, the points are in disorder, not showing an noticeable reltaionship between the number and resale price of HDB flats sold in Singapore.

(4) Differences of the resale prices of HDB flats sold in different towns.

Plot 9

This plot shows different mean resale prices from 1990 to 2019 in different towns, and sorts them by the prices.

```
data%>%
  group_by(town)%>%
  summarise(price_town=mean(prices_per_sqm))%>%
  ggplot(data=.,mapping=aes(x=reorder(town,price_town),y=price_town))+
  geom_col(fill='royalblue1')+
  coord_flip()+
  labs(x='Town',
       y='Resale price ($/sqm)',
      title='HDB Resale price in different towns in Singapore from 1990 to 2
019',
      caption='Data:Housing and Development Board,Singapore,2020')+
  theme(axis.title=element_text(size=10),axis.text.y = element_text(size=7))
```



In plot 9, we can see that the differences of resale prices among towns does not differ greatly, except for **LIM CHU KANG**(most of the land there is farmland), which is much lower than any other town. Punggol has the highest resale price of HDB sold in Singapore, and most of the towns with high prices are located in central or southern part of Singapore, like **Central Area**, **Bukit Timah**,**Bukit Merah**,**Marine Parade** and **Queenstown**.

(5) Relationship between flat type and floor area of HDB flats sold in Singapore

Plot 10

First I did data wrangling to remove the overlap data.

```
unique(data$storey_range)
```

```
##  [1] 10 TO 12 04 TO 06 07 TO 09 01 TO 03 13 TO 15 19 TO 21 16 TO 18
##  [8] 25 TO 27 22 TO 24 28 TO 30 31 TO 33 40 TO 42 37 TO 39 34 TO 36
## [15] 06 TO 10 01 TO 05 11 TO 15 16 TO 20 21 TO 25 26 TO 30 36 TO 40
## [22] 31 TO 35 46 TO 48 43 TO 45 49 TO 51
## 25 Levels: 01 TO 03 04 TO 06 07 TO 09 10 TO 12 13 TO 15 ... 49 TO 51
```

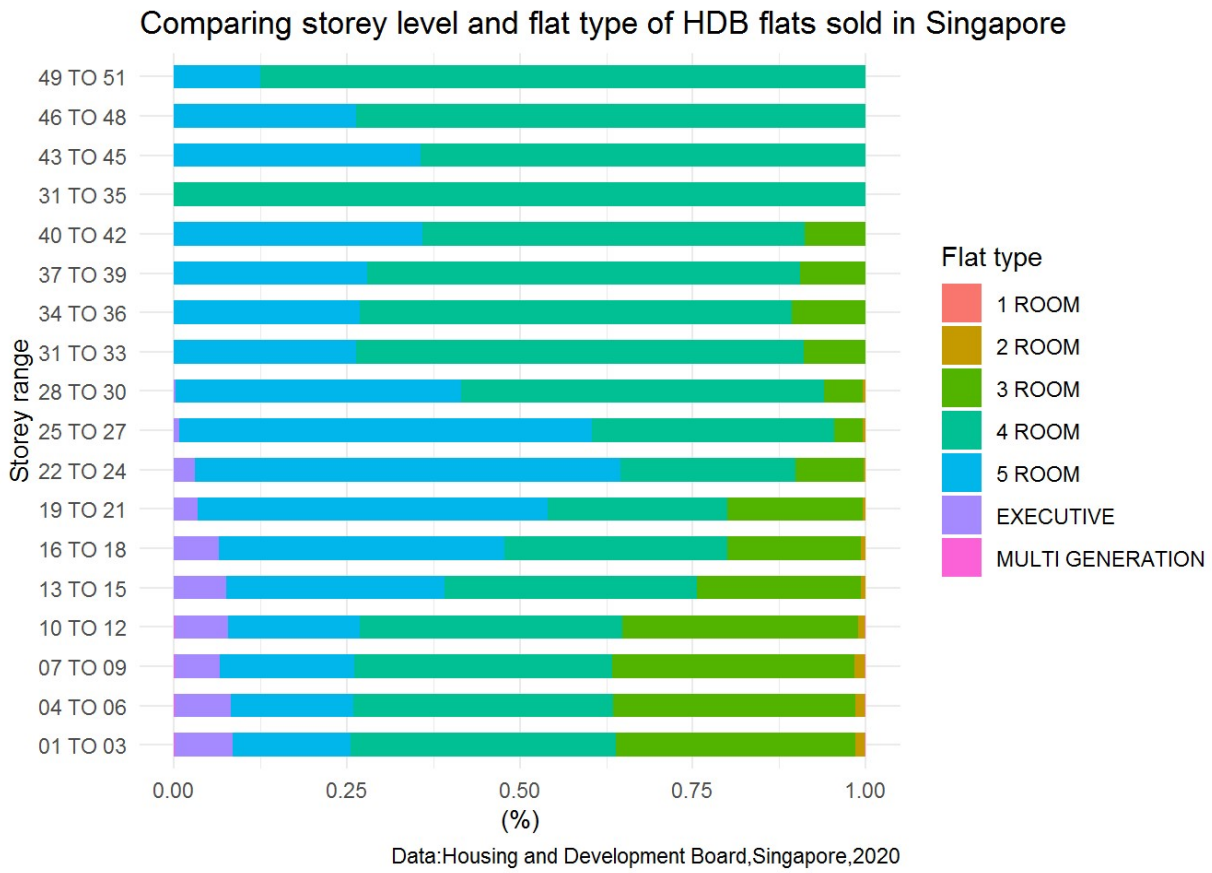
```
unique(data$flat_type)
```

```
## [1] 1 ROOM           3 ROOM           4 ROOM           5 ROOM
## [5] 2 ROOM           EXECUTIVE        MULTI GENERATION MULTI-GENERATION
## 8 Levels: 1 ROOM 2 ROOM 3 ROOM 4 ROOM 5 ROOM ... MULTI-GENERATION
```

```
data$flat_type<-recode(data$flat_type,'MULTI-GENERATION'='MULTI GENERATION')
a<-subset(data,storey_range %in%
            c('10 TO 12',
              '04 TO 06',
              '07 TO 09',
              '01 TO 03',
              '13 TO 15',
              '19 TO 21',
              '16 TO 18',
              '25 TO 27',
              '22 TO 24',
              '28 TO 30',
              '31 TO 33',
              '40 TO 42',
              '37 TO 39',
              '34 TO 36',
              '31 TO 35',
              '46 TO 48',
              '43 TO 45',
              '49 TO 51'
             ))
```

Then I used **geom_bar()** to plot three variables in one chart. The x axis is standadized value of count, showing the proportion of each kind of flat type.

```
ggplot(data=a,mapping=aes(x=storey_range,fill=flat_type))+
  geom_bar(width = 0.6,position ="fill")+
  coord_flip()+
   labs(x='Storey range',
       y='(%)',
       fill='Flat type',
      title='Comparing storey level and flat type of HDB flats sold in Singa
pore',
      caption='Data:Housing and Development Board,Singapore,2020')+
  theme_minimal()+
  theme(axis.title=element_text(size=10))
```



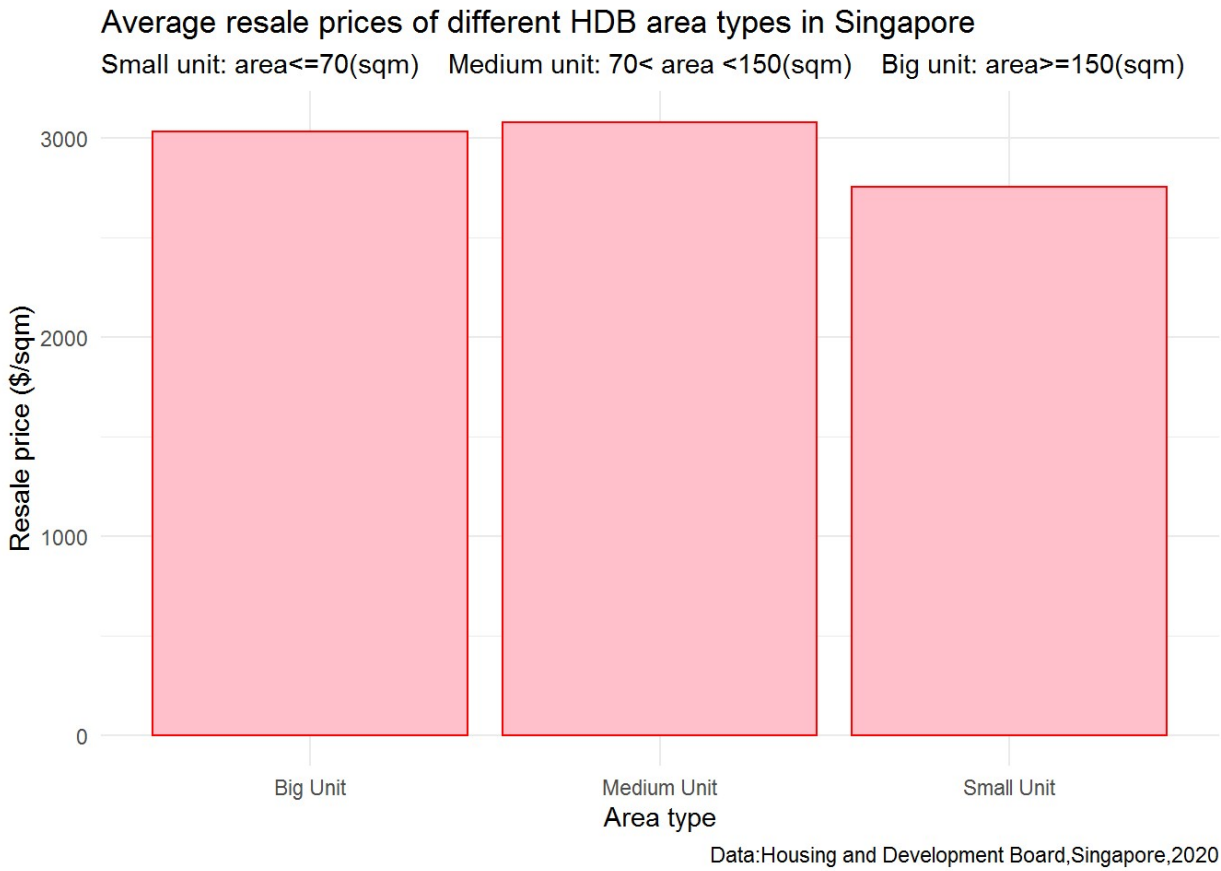
Plot 10 indicates : More 4-room flats are sold in high levels (more than 28 to 30). More 3-room flats are sold in low levels (less than 10 to 12). More 5-room flats are sold in middle levels (13 to 15 - 22 to 27). Most of the executive flats are sold in levels under 25.

(6) Relationship between flat area and resale price of HDB flats sold in Singapore

Plot 11

To figure out if there is a relationship between flat area and resale price ($/sqm), I divided the values of flat area into three groups: samll unit(area<=70 sqm),Medium unit(70<area<150 sqm),and Big unit(area>=150 sqm). Then calculated the mean resale prices of these three groups and plot it in below.

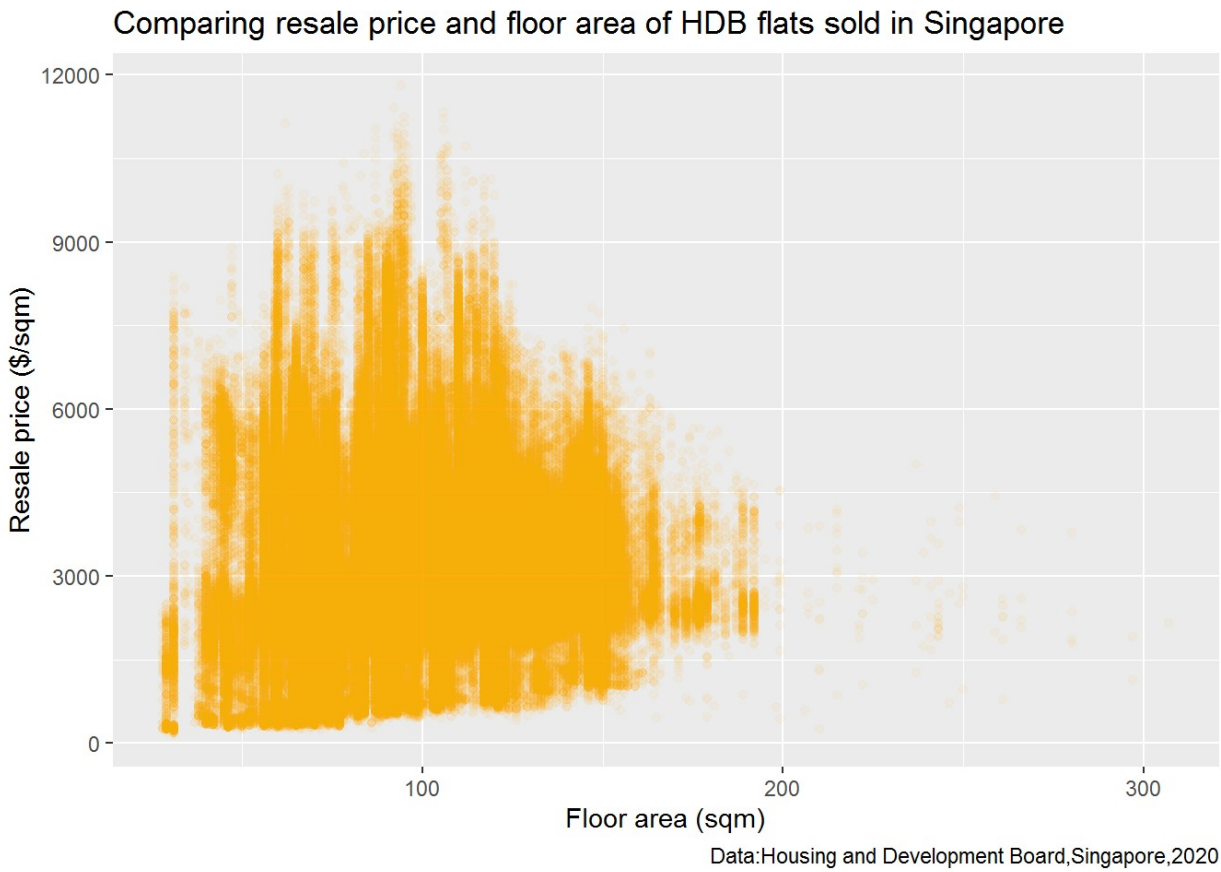
```
data%>%
  mutate(area_type= case_when(
    floor_area_sqm<=70 ~'Small Unit',
    floor_area_sqm>70 &floor_area_sqm<150~ 'Medium Unit',
    floor_area_sqm>=150~ 'Big Unit'
  ))%>%
  group_by(area_type)%>%
  summarise(area_type_price=mean(prices_per_sqm))%>%
  ggplot(data=.,mapping=aes(x=area_type,y=area_type_price))+
  geom_col(color='red',fill='pink')+
  labs(x='Area type',
       y='Resale price ($/sqm)',
       title='Average resale prices of different HDB area types in Singapor
e',
       subtitle='Small unit: area<=70(sqm)    Medium unit: 70< area <150(sq
m)    Big unit: area>=150(sqm)',
       caption='Data:Housing and Development Board,Singapore,2020')+
       theme_minimal()
```



Plot 11 shows that **big unit** and **medium unit** have almost the same resale price ($/sqm), and the price of **small unit** is a little lower than the other two groups.

Plot 12

```
  ggplot(data=data,mapping=aes(x=floor_area_sqm,y=prices_per_sqm))+
    geom_point(color='orange',alpha=0.05)+
 labs(y='Resale price ($/sqm)',
       x='Floor area (sqm)',
       title='Comparing resale price and floor area of HDB flats sold in Sin
gapore',
       caption='Data:Housing and Development Board,Singapore,2020')
```



In this scatter plot, most of the points are concentrating in the bottom left corner, showing that flats with small area and low resale price ($/sqm) are sold the most, and there is not a noticeable linear relationship between floor area and resale price.

Task 2 RECREATING A NEW PLOT TYPE

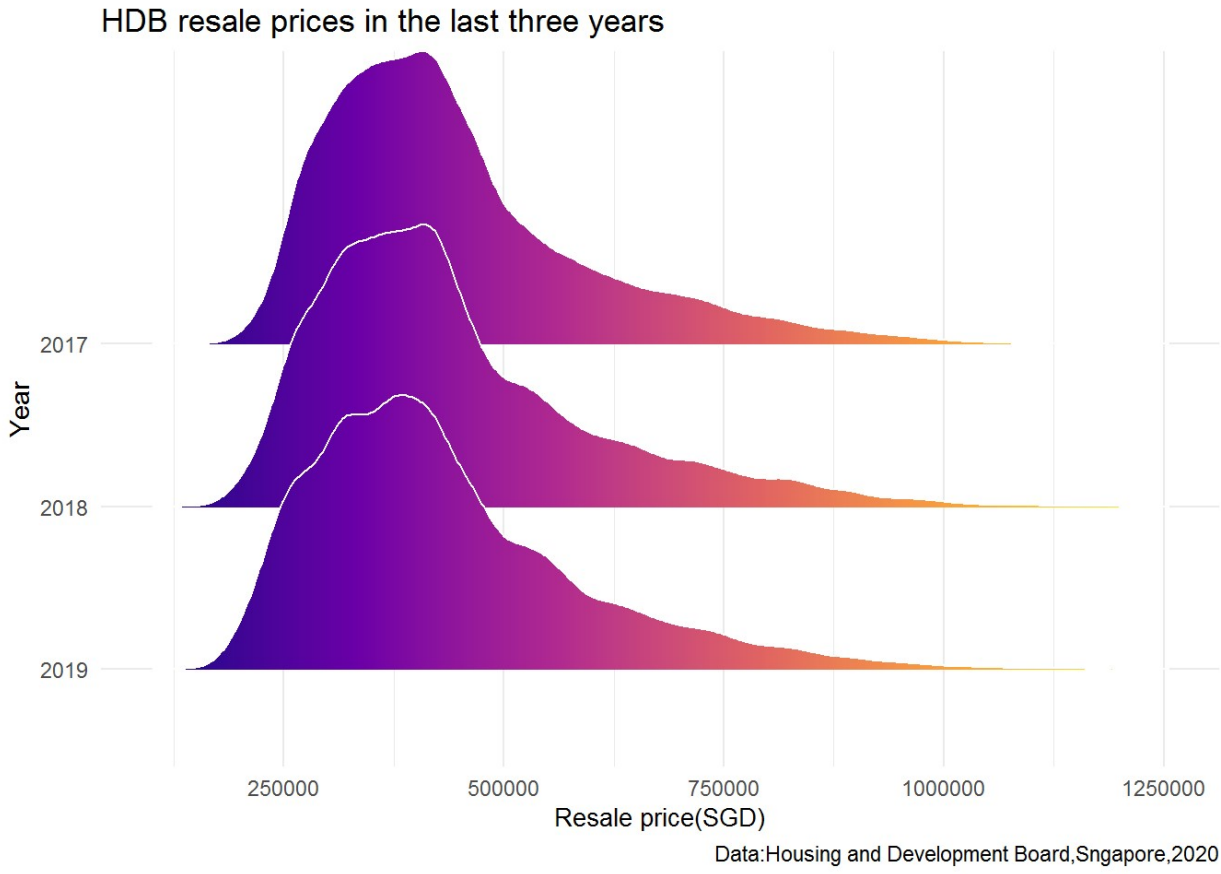
2.1 Plots and Analysis

Plot 1

```
task2<-data%>%
  filter(year%in%c('2019','2018','2017'))

ggplot(data=task2,mapping=aes(x=resale_price,y=reorder(year,resale_price),fi
ll=stat(x)))+
  geom_density_ridges_gradient(color='white',scale=1.8)+
  scale_fill_viridis_c(name='Resale_price',option='C')+
  theme_minimal()+
  labs(x='Resale price(SGD)',
       y='Year',
       title='HDB resale prices in the last three years',
       caption='Data:Housing and Development Board,Sngapore,2020')+
  theme(axis.title.x=element_text(size=10))+
  theme(legend.position="none")
```

```
## Picking joint bandwidth of 16700
```

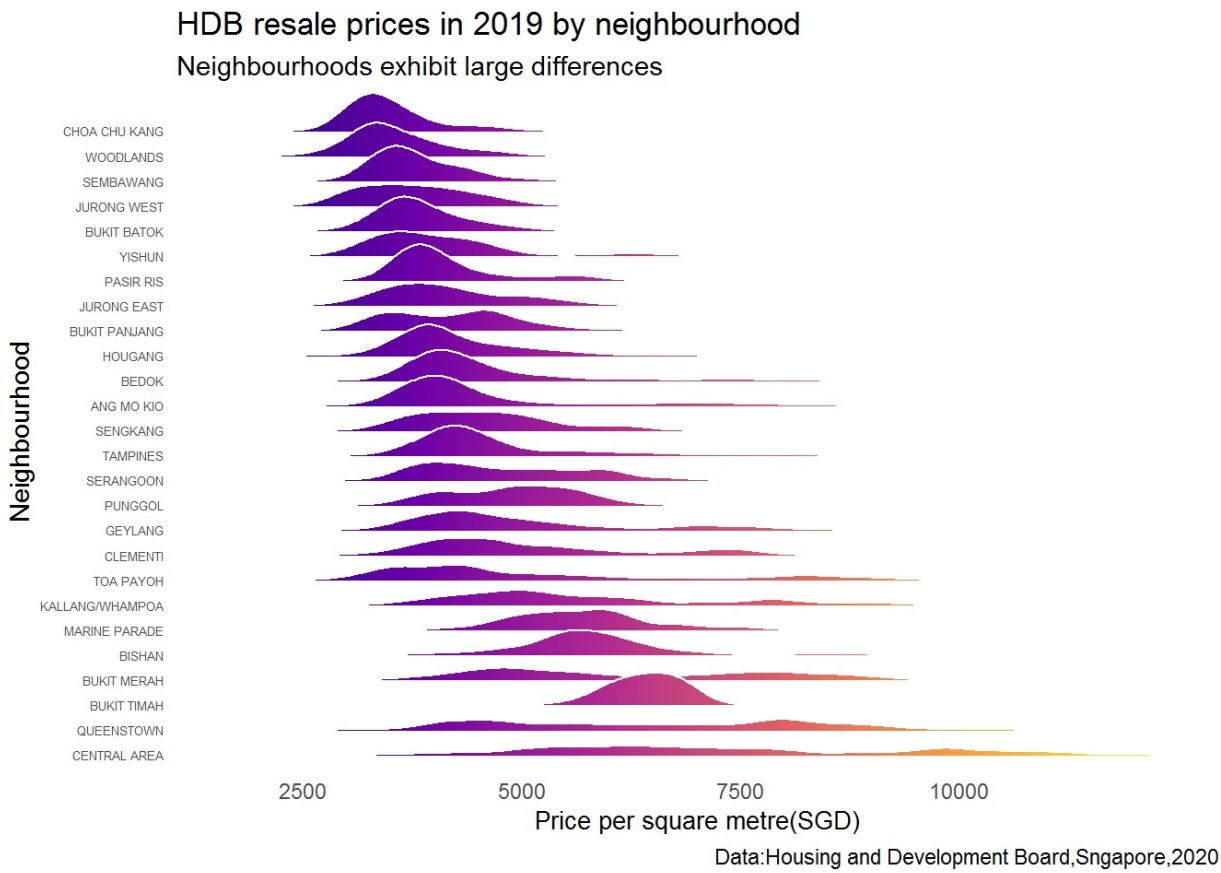


From this plot,we can see that the distribution of resale price for year 2017, 2018 and 2019 are very similar. Their peaks are all around 406000 (SGD), and most of the flats with a price lower than 500000 (SGD). As for the max value of resale price, year 2019 has the highest max price, while year 2018 has the lowest.

Plot 2

```
task2_2<-data%>%
  filter(year=='2019')
ggplot(data=task2_2,mapping=aes(x=prices_per_sqm,y=reorder(town,-prices_per_
sqm),fill=stat(x)))+
  geom_density_ridges_gradient(color='white',scale=1.5)+
  scale_fill_viridis_c(name='Resale_price',option='C')+
  theme_minimal()+
  labs(x='Price per square metre(SGD)',
       y='Neighbourhood',
       title='HDB resale prices in 2019 by neighbourhood',
       subtitle='Neighbourhoods exhibit large differences',
       caption='Data:Housing and Development Board,Sngapore,2020')+
  theme(axis.title.x=element_text(size=10),axis.text.y=element_text(size=5))
+
  theme(legend.position="none")+
   scale_x_continuous(breaks=c(2500,5000,7500,10000),labels=c('2500','500
0','7500','10000'))+
  theme(panel.grid = element_blank())
```

```
## Picking joint bandwidth of 205
```



From this plot, we can see a general trend that: neighbourhood with lower prices (SGD/sqm) have more concentrated distribution (the height of the curve is higher),except for Bishan and Bukit Timah, which have relative highrt prices (SGD/sqm) but concentrated distribution.

Task 3 YOUR DATA, YOUR ANALYSIS

3.1 Preparation

First,I downloaed the data *Crimes in Boston* from Kaggle (https://www.kaggle.com/AnalyzeBoston/crimes-in-boston), and imported into R Studio.

```
crime <- read.csv("task3 data/crimes-in-boston/crime.csv")
```

Then, install “ggmap” package for ploting spatial information on a map.

```
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
##
## Attaching package: 'ggmap'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     crime
```

3.2 Data Wrangling

First I used **unique** to check how many different types of crime are in this dataset, and found there are 67 types, which is too many for my analysis.

```
unique(crime$OFFENSE_CODE_GROUP)
```

```
##  [1] Larceny
##  [2] Vandalism
##  [3] Towed
##  [4] Investigate Property
##  [5] Motor Vehicle Accident Response
##  [6] Auto Theft
##  [7] Verbal Disputes
##  [8] Robbery
##  [9] Fire Related Reports
## [10] Other
## [11] Property Lost
## [12] Medical Assistance
## [13] Assembly or Gathering Violations
## [14] Larceny From Motor Vehicle
## [15] Residential Burglary
## [16] Simple Assault
## [17] Restraining Order Violations
## [18] Violations
## [19] Harassment
## [20] Ballistics
## [21] Property Found
## [22] Police Service Incidents
## [23] Drug Violation
## [24] Warrant Arrests
## [25] Disorderly Conduct
## [26] Property Related Damage
## [27] Missing Person Reported
## [28] Investigate Person
## [29] Fraud
## [30] Aggravated Assault
## [31] License Plate Related Incidents
## [32] Firearm Violations
## [33] Other Burglary
## [34] Arson
## [35] Bomb Hoax
## [36] Harbor Related Incidents
## [37] Counterfeiting
## [38] Liquor Violation
## [39] Firearm Discovery
## [40] Landlord/Tenant Disputes
## [41] Missing Person Located
## [42] Auto Theft Recovery
## [43] Service
## [44] Operating Under the Influence
## [45] Confidence Games
## [46] Search Warrants
## [47] License Violation
## [48] Commercial Burglary
## [49] HOME INVASION
## [50] Recovered Stolen Property
## [51] Offenses Against Child / Family
## [52] Prostitution
```

```
## [53] Evading Fare
## [54] Prisoner Related Incidents
## [55] Homicide
## [56] Embezzlement
## [57] Explosives
## [58] Criminal Harassment
## [59] Phone Call Complaints
## [60] Aircraft
## [61] Biological Threat
## [62] Manslaughter
## [63] Gambling
## [64] INVESTIGATE PERSON
## [65] HUMAN TRAFFICKING
## [66] HUMAN TRAFFICKING - INVOLUNTARY SERVITUDE
## [67] Burglary - No Property Taken
## 67 Levels: Aggravated Assault Aircraft ... Warrant Arrests
```

To make the analysis easier, i reocode the **OFFENSE_CODE_GROUP** to divide the crimes into eleven board types (acoording to Definition of Crime Categories (https://campusadvisories.gwu.edu/definition-crime-categories)), and removed some of the crimes that is not included in these categories.

```
crime$category<-recode(crime$OFFENSE_CODE_GROUP,
               'Residential'= 'Burglary',
               'Other Burglary'='Burglary',
               'Commercial Buglary'='Burglary',
               'HOME INVASION'='Burglary',
               'Burglary - No Property Taken'='Burglary',
               'Property Related Damage'='Destruction/Damage/Vandalism of Pr
operty',
               'Offenses Against Child / Family '='Domestic Violence',
               'Drug Violation'='Drug Abuse Violations',
               'Liquor Violation'='Liquor Law Violations',
               'Mauslaughter'='Mauslaughter by Negligence',
               'Larceny From Motor Vehicle'='Motor Vehicle Theft',
               'Auto Theft'='Motor Vehicle Theft',
               'Homicide'='Murder and Non-negligent Mauslaughter',
               'Criminal Harassment'='Harassment',
               'Firearm Violations'='Weapons Law Violations',
               'Confidence Games'='Fraud')

crime2<-crime%>%
  filter(category %in% c('Aggravated Assault','Arson','Burglary','Destructio
n/Damage/Vandalism of Property','Property Related Damage','Larceny','Robber
y','Simple Assault','Harassment','Fraud','Bomb Hoax','Mauslaughter by Neglig
ence','Motor Vehicle Theft','Murder and Non-negligent Mauslaughter'))

crime2$type<-recode(crime2$category,
                    'Aggravated Assault'='Assault',
                    'Simple Assault'='Assault',
                    'Property Related Damage'='Destruction/Damage/Vandalism
of Property')
```

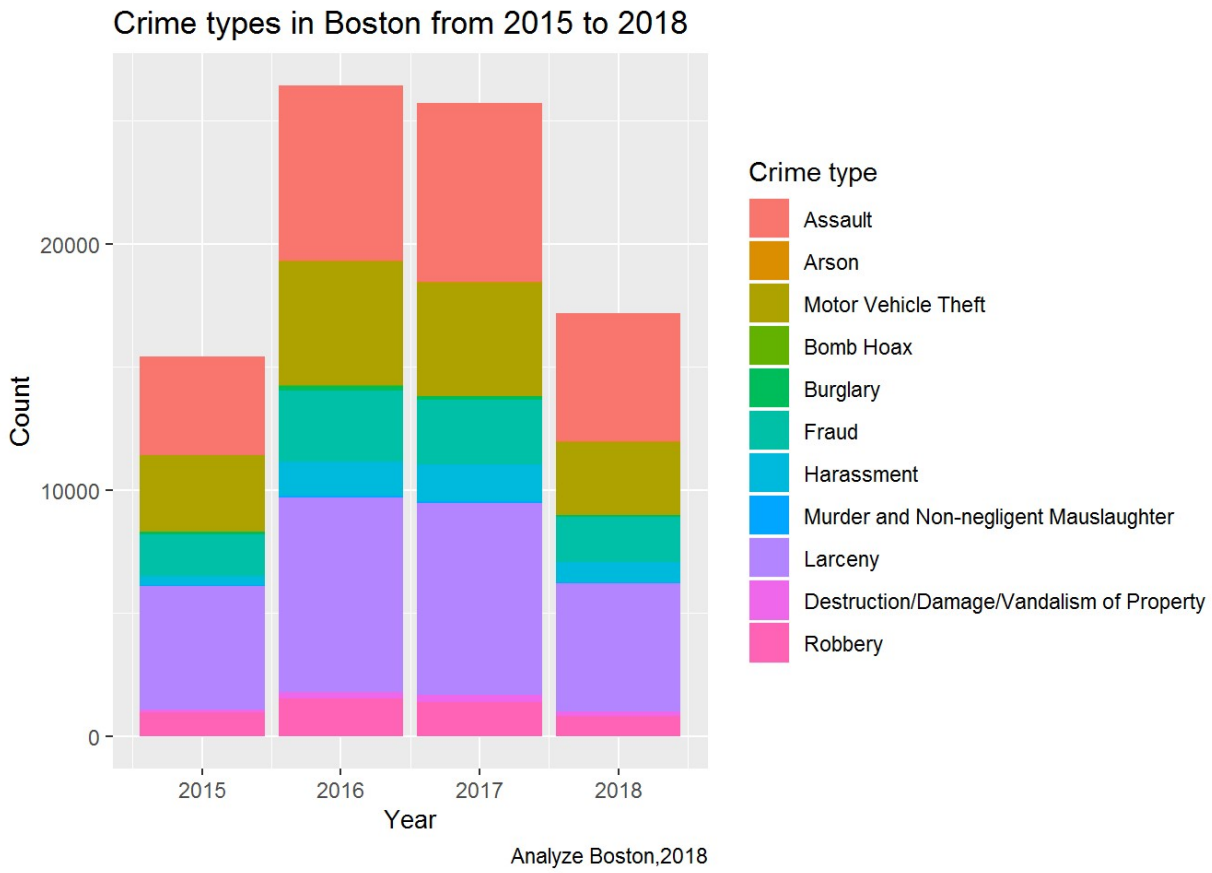
3.3 Plots and Analysis

(1) Crime distribution by time (year, month, day of week, hour of day)

Plot 1 (by year)

```
crime2<-crime2%>%
  mutate(day=case_when(
    DAY_OF_WEEK=='Monday'~'1',
    DAY_OF_WEEK=='Tuesday'~'2',
    DAY_OF_WEEK=='Wednesday'~'3',
    DAY_OF_WEEK=='Thursday'~'4',
    DAY_OF_WEEK=='Friday'~'5',
    DAY_OF_WEEK=='Saturday'~'6',
    DAY_OF_WEEK=='Sunday'~'7'
  ))

ggplot()+
  geom_bar(data=crime2,mapping=aes(x=YEAR,fill=type),width = 0.9,position
="stack")+
  labs(x='Year',
       y='Count',
       fill='Crime type',
       title='Crime types in Boston from 2015 to 2018',
       caption='Analyze Boston,2018')+
  theme(axis.title.x=element_text(size=10))
```

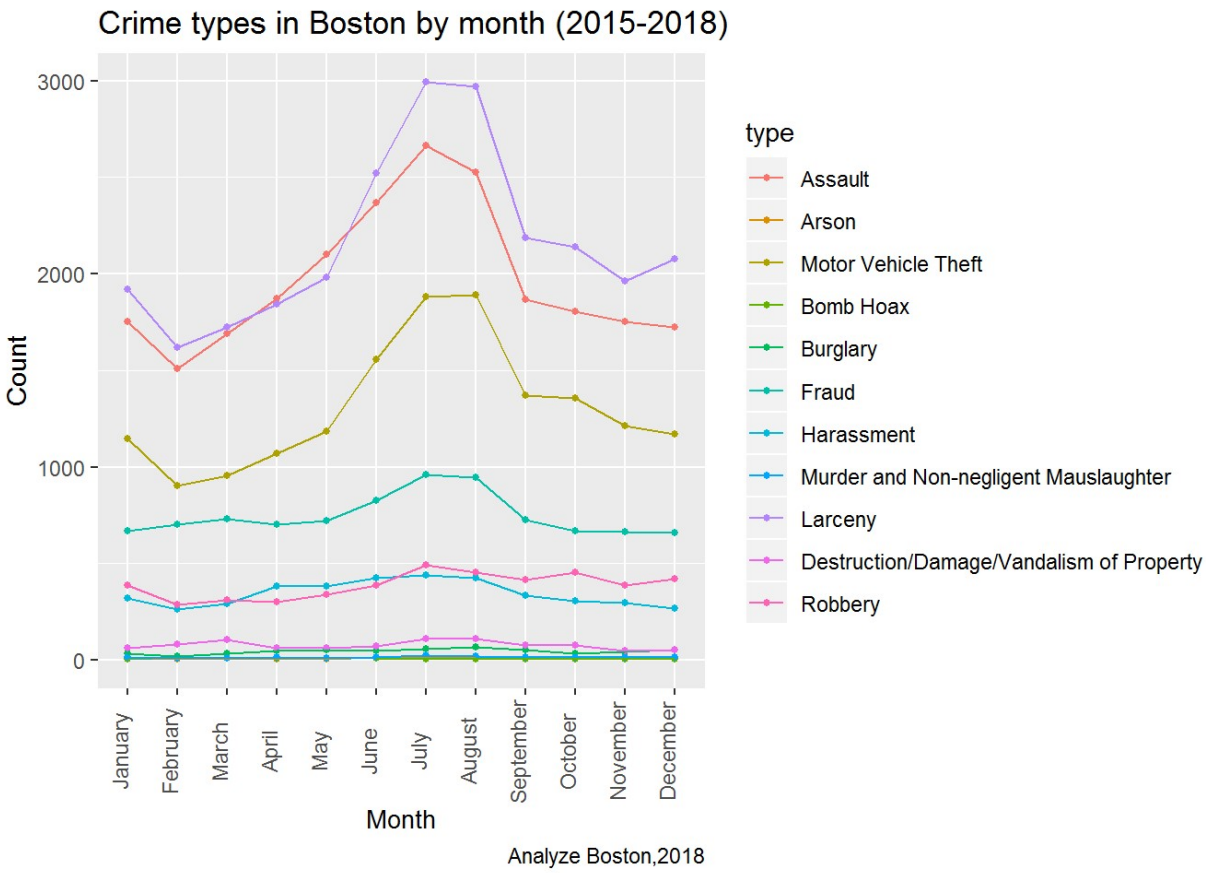


Plot 1 shows that there are significantly more crimes in 2016 and 2017 than in 2015 and 2018 in Boston. The count of different types of crime differs greatly. Larceny, assault and motor vehicle theft are the three most common crime types. Serious crimes like Murder and non-negligent mauslaughter and Bomb hoax are not very often to happen.

Plot 2 (by month)

```
crime2%>%
  group_by(MONTH,type)%>%
  summarise(b=n())%>%
  mutate(Month=case_when(
    MONTH=='1'~'January',
    MONTH=='2'~'February',
    MONTH=='3'~'March',
    MONTH=='4'~'April',
    MONTH=='5'~'May',
    MONTH=='6'~'June',
    MONTH=='7'~'July',
    MONTH=='8'~'August',
    MONTH=='9'~'September',
    MONTH=='10'~'October',
    MONTH=='11'~'November',
    MONTH=='12'~'December'
  ))%>%
  ggplot(data=.,mapping=aes(x=reorder(Month,MONTH),y=b,color=type,group=typ
e))+
  geom_point(size=1)+
  geom_line(line=1)+
  theme(axis.title.x=element_text(size=10),axis.text.x = element_text(angle
= 90,vjust=0))+
  labs(x='Month',
       y='Count',
       fill='Crime type',
       title='Crime types in Boston by month (2015-2018)',
       caption='Analyze Boston,2018')
```

```
## Warning: Ignoring unknown parameters: line
```

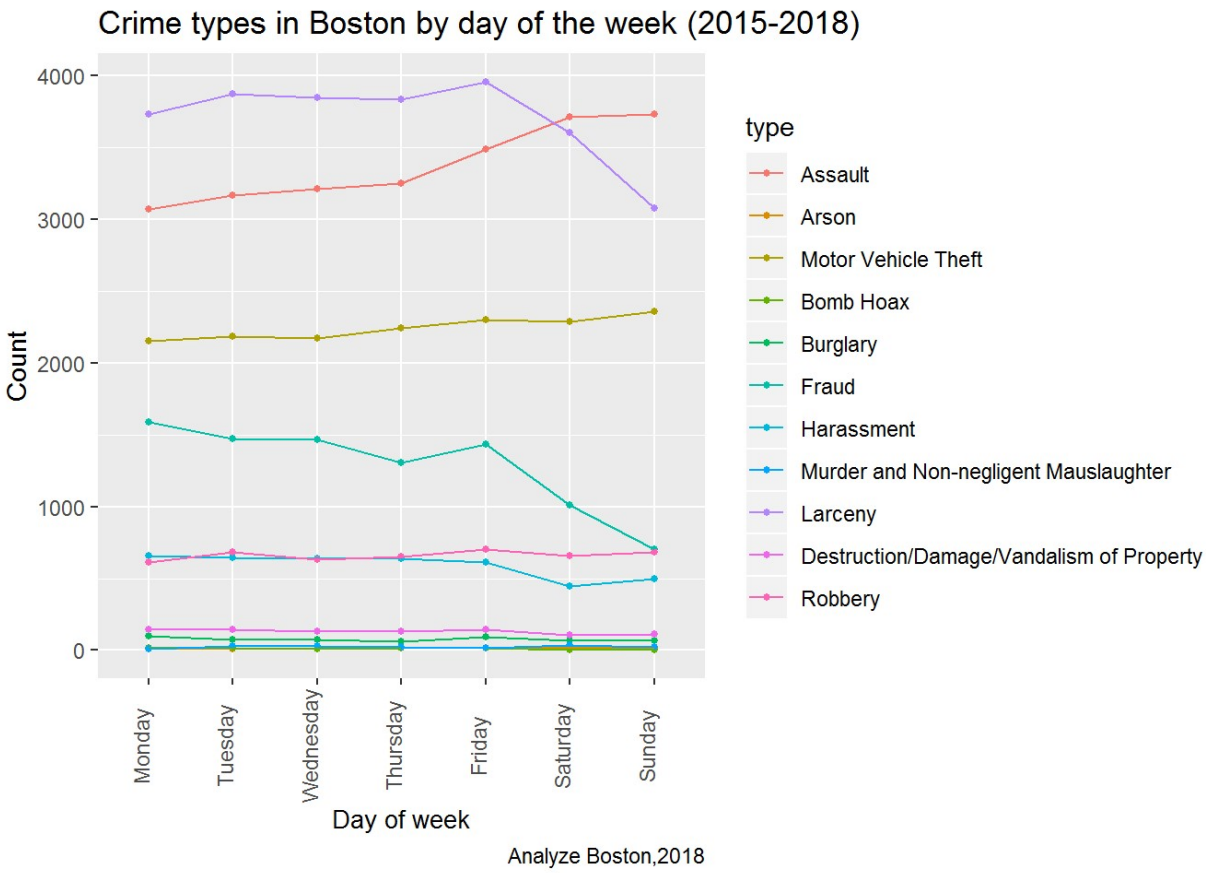


Plot 2 shows that the July and August (summer) have most crimes. Most of the crime types show a general upward trend from Jauary to August, following by a downward trend till December.

Plot 3

```
crime2%>%
  group_by(day,type)%>%
  summarise(a=n())%>%
  ggplot(data=.,mapping=aes(x=day,y=a,color=type,group=type))+
  geom_point(size=1)+
  geom_line(line=1)+
  theme(axis.title.x=element_text(size=10),axis.text.x = element_text(angle
= 90,vjust=0))+
  labs(x='Day of week',
       y='Count',
       fill='Crime type',
       title='Crime types in Boston by day of the week (2015-2018)',
       caption='Analyze Boston,2018')+
  scale_x_discrete(breaks=c(1,2,3,4,5,6,7),label=c('Monday','Tuesday','Wedne
sday','Thursday','Friday','Saturday','Sunday'))
```

```
## Warning: Ignoring unknown parameters: line
```

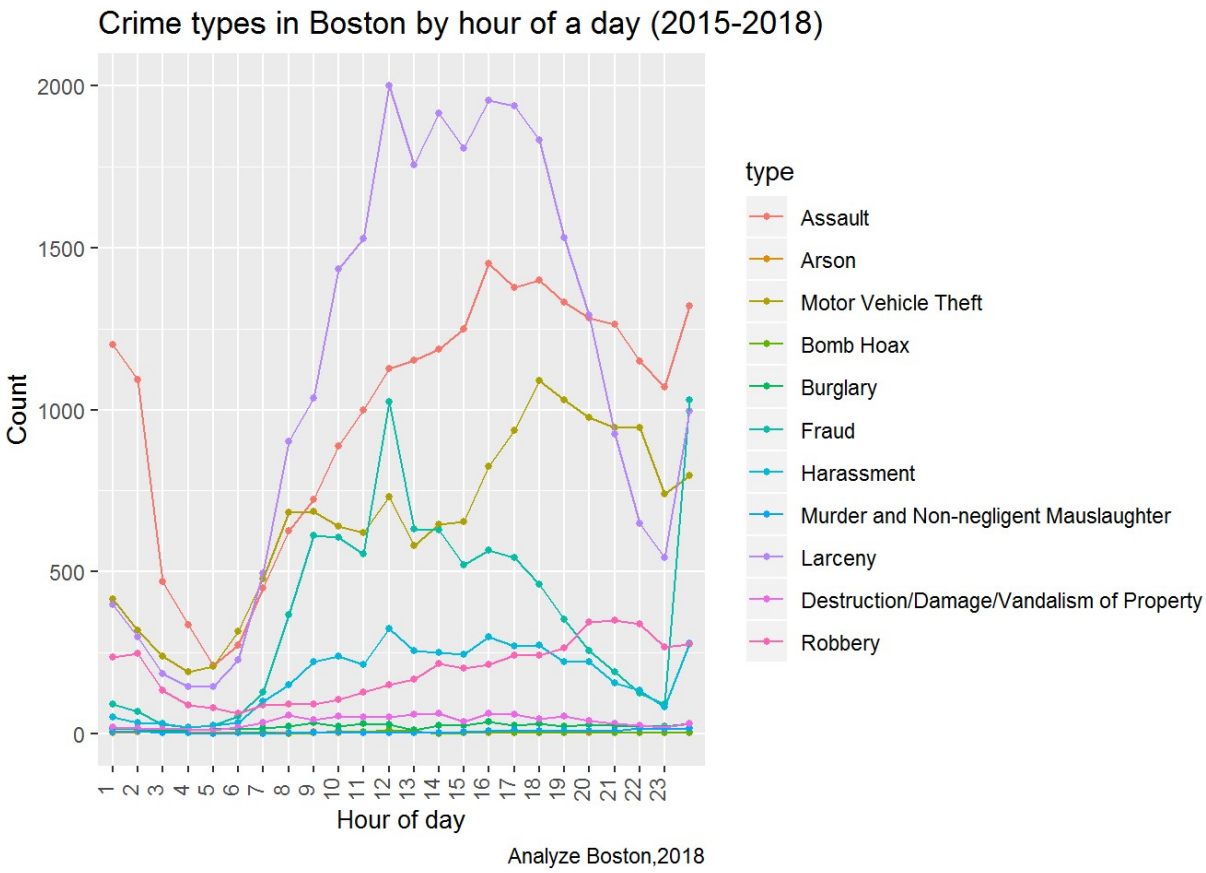


Plot 3 shows that in weekdays, the count of different types of crime stay stable. In weekend,Larceny, fraud and harassment show obvious downward trends, while assult shows a noticeable upward trend.

Plot 4

```
crime2%>%
  group_by(HOUR,type)%>%
  summarise(d=n())%>%
  mutate(Hour=case_when(
    HOUR=='1'~'1',
    HOUR=='2'~'2',
    HOUR=='3'~'3',
    HOUR=='4'~'4',
    HOUR=='5'~'5',
    HOUR=='6'~'6',
    HOUR=='7'~'7',
    HOUR=='8'~'8',
    HOUR=='9'~'9',
    HOUR=='10'~'10',
    HOUR=='11'~'11',
    HOUR=='12'~'12',
    HOUR=='13'~'13',
    HOUR=='14'~'14',
    HOUR=='15'~'15',
    HOUR=='16'~'16',
    HOUR=='17'~'17',
    HOUR=='18'~'18',
    HOUR=='19'~'19',
    HOUR=='20'~'20',
    HOUR=='21'~'21',
    HOUR=='22'~'22',
    HOUR=='23'~'23',
  ))%>%
  ggplot(data=.,mapping=aes(x=reorder(Hour,HOUR),y=d,color=type,group=type))
+
  geom_point(size=1)+
  geom_line(line=1)+
  theme(axis.title.x=element_text(size=10),axis.text.x = element_text(angle
= 90,vjust=0))+
  labs(x='Hour of day',
       y='Count',
       fill='Crime type',
       title='Crime types in Boston by hour of a day (2015-2018)',
       caption='Analyze Boston,2018')+
 scale_x_discrete(breaks=c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,1
9,20,21,22,23),label=c('0','1','2','3','4','5','6','7','8','9','10','11','1
2','13','14','15','16','17','18','19','20','21','22','23'))
```

```
## Warning: Ignoring unknown parameters: line
```



In plot 4, We can see drastic changes of the count of different crime types in different hours. Generally, 11:00 to 19:00 is the period that most of the crimes happen. The numbers of most of the crimes decline after sunset, but raise again during 23:00 to 0:00. And the number of Larceny decreases most drasticlly after 18:00.

(2) Crime distribution by location

Plot 5

I made a scatter plot on Google map to show the spatial distribution of crimes in Boston.

```
require(ggmap)
register_google(key = "AIzaSyBlXzgWHU7ENAbA-yWE_Kcn4pydQkiRisk")
map.center <- geocode("Grove Hall,Boston")
```

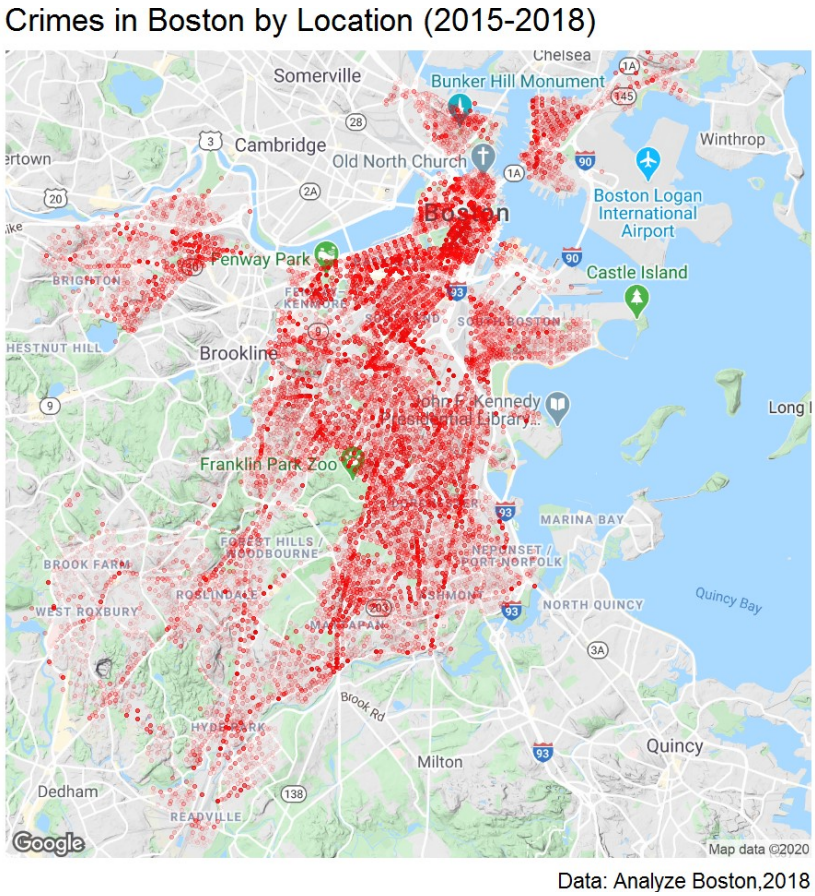
```
## Source : https://maps.googleapis.com/maps/api/geocode/json?address=Grove+
Hall,Boston&key=xxx-yWE_Kcn4pydQkiRisk
```

```
Bos_map <- qmap(c(lon=map.center$lon, lat=map.center$lat), zoom=12)
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=42.311209,
-71.074495&zoom=12&size=640x640&scale=2&maptype=terrain&language=en-EN&key=x
xx-yWE_Kcn4pydQkiRisk
```

```
g <- Bos_map +
  geom_point(aes(x=Long, y=Lat), data=crime2, size=0.5, alpha=0.04, color="r
ed2") +
  ggtitle("Crimes in Boston by Location (2015-2018)")+
  labs(caption='Data: Analyze Boston,2018')
g
```

```
## Warning: Removed 3957 rows containing missing values (geom_point).
```



This plot shows that the most densely concentrated area of crimes is the CBD in northern Boston, which has dense road network and high population density. And the number of crimes decreases from CBD to peripheral area. The region near Dedham has least crime number.