# # Prices of Housing Development Board (HDB) Appartments

In [96]:

```
# Load dataset
import csv
path = "C:/Users/justm/Documents/Python Training/resale-flat-prices-based-on-regi
f = open(path)
all_lines = csv.reader(f, delimiter = ',')
header=next(all_lines)
header
next(all_lines)
```

Out[96]:

```
['2017-01',
 'ANG MO KIO',
 '2 ROOM',
 '406',
 'ANG MO KIO AVE 10',
 '10 TO 12',
 '44',
 'Improved',
 '1979',
 '61 years 04 months',
 '232000']
```

In [97]:

```

# Data cleaning
import time

dataset = []

for line in all_lines:
    d = dict(zip(header, line))
    d['month_structured_obj'] = time.strptime(d['month'],'%Y-%m')
    d['month_number'] = time.mktime(d['month_structured_obj'])
    d['floor_area_sqm'] = float(d['floor_area_sqm'])
    d['lease_commence_date'] = int(d['lease_commence_date'])
    d['resale_price'] = float(d['resale_price'])
    d['remaining_lease_years'] = float(d['remaining_lease'].split(' year')[0])
    if d['remaining_lease'].find('month')>0 :
        d['remaining_lease_months'] = float(d['remaining_lease'].split(' year')[
    else:
        d['remaining_lease_months'] =0.0
    d['remaining_lease_ttl_months']=d['remaining_lease_years']*12+d['remaining_le
    dataset.append(d)

d
```

Out[97]:

```
{'month': '2020-03',
 'town': 'YISHUN',
 'flat_type': 'EXECUTIVE',
 'block': '827',
 'street_name': 'YISHUN ST 81',
 'storey_range': '01 TO 03',
 'floor_area_sqm': 145.0,
 'flat_model': 'Maisonette',
 'lease_commence_date': 1987,
 'remaining_lease': '66 years 07 months',
 'resale_price': 660000.0,
 'month_structured_obj': time.struct_time(tm_year=2020, tm_mon=3, tm_mday=1, tm
_hour=0, tm_min=0, tm_sec=0, tm_wday=6, tm_yday=61, tm_isdst=-1),
 'month_number': 1582992000.0,
 'remaining_lease_years': 66.0,
 'remaining_lease_months': 7.0,
 'remaining_lease_ttl_months': 799.0}
```

In [98]:

```
# What is the total number of entries in the dataset?
len(dataset)
```

Out[98]:

```
70102
```

In [99]:

```
# What is the average resale price?
import numpy
numpy.mean([d['resale_price'] for d in dataset])
```

Out[99]:

```
438402.7636094549
```

In [129]:

```
# What is the average resale price by flat type?
import numpy
from collections import defaultdict
import pandas as pd

nResale_price_cnt=defaultdict(int)
nResale_price_flattype=defaultdict(int)

for d in dataset:
    nResale_price_cnt[d['flat_type']]+=1
    nResale_price_flattype[d['flat_type']]+=d['resale_price']


df_avg_Resale_price_flattype=pd.DataFrame(nResale_price_flattype.values())/pd.Dat
df_avg_Resale_price_flattype['flat_type']=nResale_price_flattype.keys()
df_avg_Resale_price_flattype.columns = ['avg_resale_price', 'flat_type']
df_avg_Resale_price_flattype=df_avg_Resale_price_flattype.sort_values('flat_type'
df_avg_Resale_price_flattype
```

Out[129]:

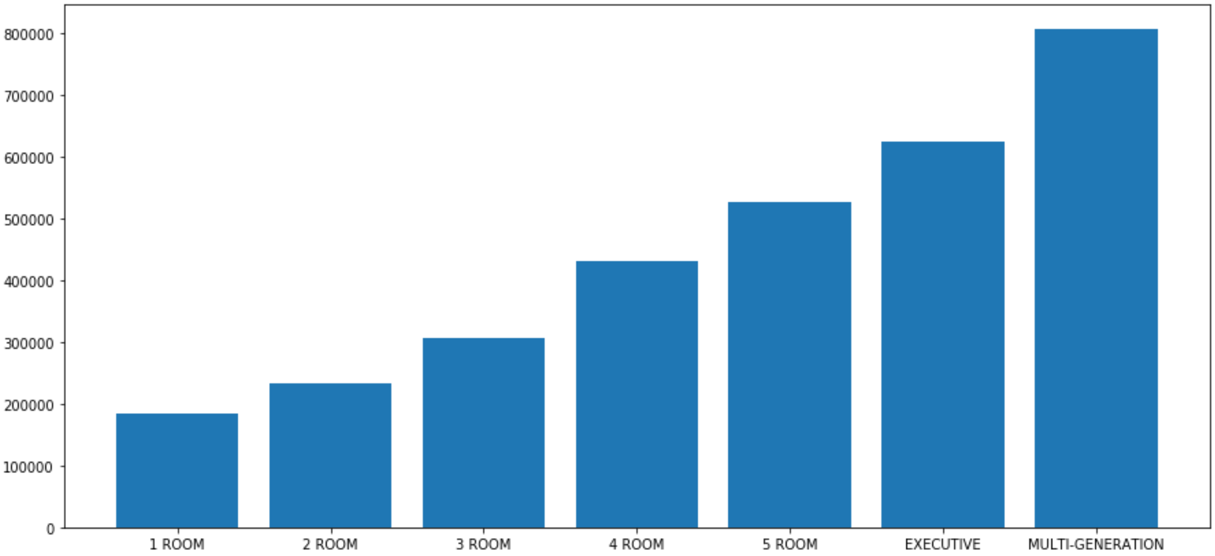
| | avg_resale_price | flat_type |
|---|---|---|
| **5** | 183789.625000 | 1 ROOM |
| **4** | 233366.089423 | 2 ROOM |
| **0** | 307356.564567 | 3 ROOM |
| **1** | 432210.852958 | 4 ROOM |
| **2** | 528232.292763 | 5 ROOM |
| **3** | 625316.397820 | EXECUTIVE |
| **6** | 806804.606061 | MULTI-GENERATION |

In [130]:

```
import matplotlib.pyplot as plt
from matplotlib import colors
import numpy

plt.figure(figsize=(15, 7))

plt.bar(df_avg_Resale_price_flattype['flat_type'], df_avg_Resale_price_flattype['
```

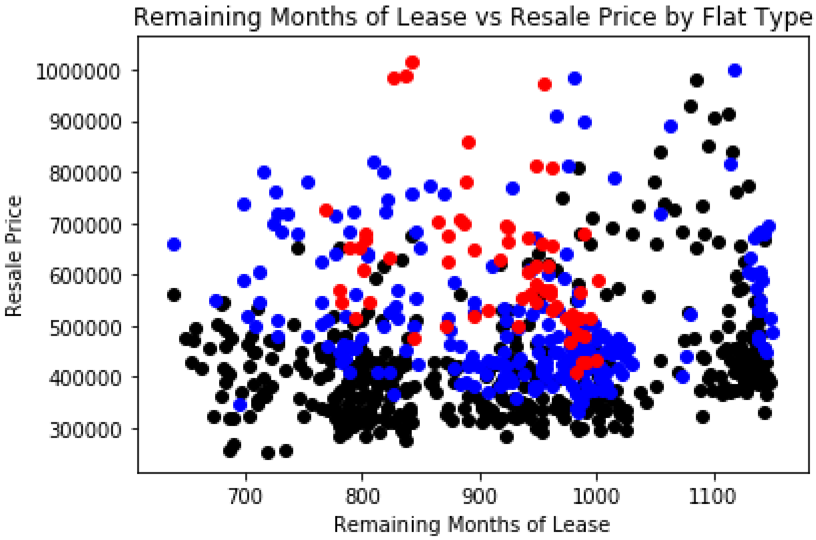
Out[130]:

```
<BarContainer object of 7 artists>
```

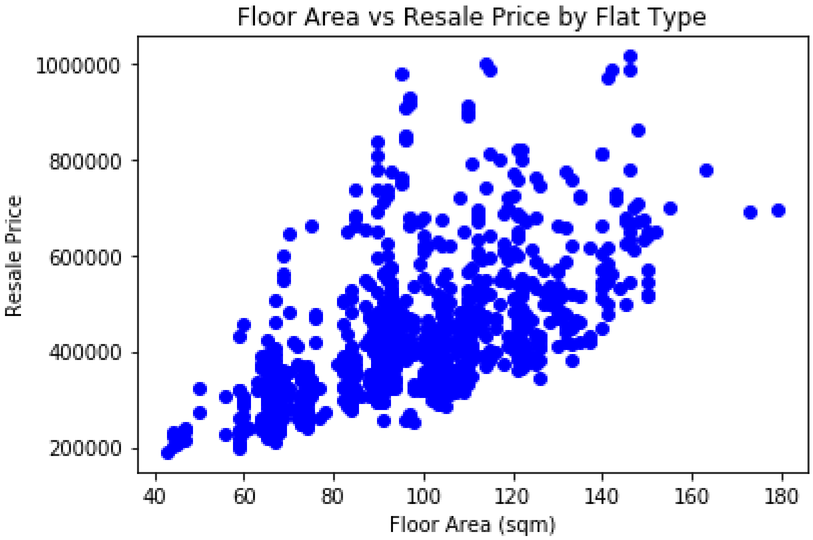


In [142]:

```
# Scatter plot of remaining lease vs resale price (sample because too many datapo
import random

dataset_sample=random.sample(dataset,1000)

remaining_lease_ttl_months_4room = [d['remaining_lease_ttl_months'] for d in data
resale_price_4room = [d['resale_price'] for d in dataset_sample if d['flat_type']
plt.scatter(remaining_lease_ttl_months_4room, resale_price_4room, c='black')

remaining_lease_ttl_months_5room = [d['remaining_lease_ttl_months'] for d in data
resale_price_5room = [d['resale_price'] for d in dataset_sample if d['flat_type']
plt.scatter(remaining_lease_ttl_months_5room, resale_price_5room, c='blue')

remaining_lease_ttl_months_EXECUTIVE = [d['remaining_lease_ttl_months'] for d in
resale_price_EXECUTIVE = [d['resale_price'] for d in dataset_sample if d['flat_ty
plt.scatter(remaining_lease_ttl_months_EXECUTIVE, resale_price_EXECUTIVE, c='red'

plt.gca().set(title='Remaining Months of Lease vs Resale Price by Flat Type', yla
plt.show()
```

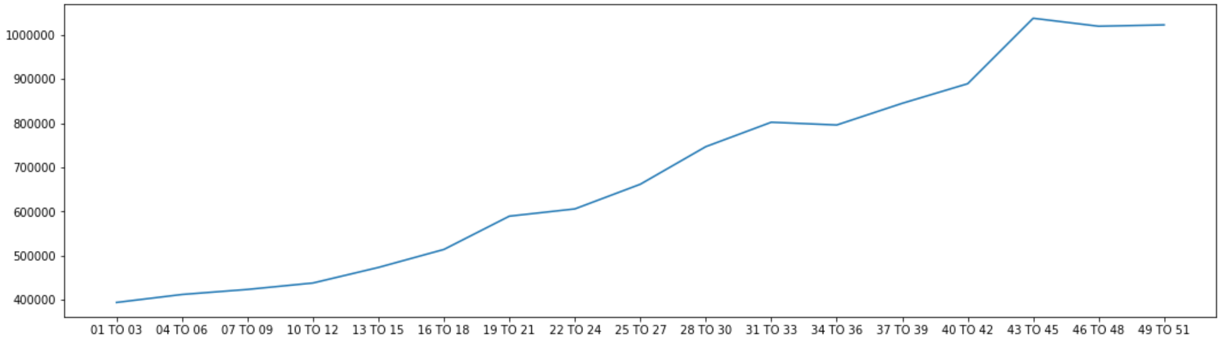


In [143]:
```
floor_area = [d['floor_area_sqm'] for d in dataset_sample]
resale_price = [d['resale_price'] for d in dataset_sample]

plt.scatter(floor_area, resale_price, c='blue')

plt.gca().set(title='Floor Area vs Resale Price by Flat Type', ylabel='Resale Pri
plt.show()
```



In [149]:

```
# What is the average resale price by storey range?
import numpy
from collections import defaultdict
import pandas as pd

nResale_price_cnt=defaultdict(int)
nResale_price_storey_range=defaultdict(int)

for d in dataset:
    nResale_price_cnt[d['storey_range']]+=1
    nResale_price_storey_range[d['storey_range']]+=d['resale_price']


df_avg_Resale_price_storey_range=pd.DataFrame(nResale_price_storey_range.values()
df_avg_Resale_price_storey_range['storey_range']=nResale_price_storey_range.keys(
df_avg_Resale_price_storey_range.columns = ['avg_resale_price', 'storey_range']
df_avg_Resale_price_storey_range=df_avg_Resale_price_storey_range.sort_values('st
#df_avg_Resale_price_storey_range

plt.figure(figsize=(18,5))
plt.plot(df_avg_Resale_price_storey_range['storey_range'], df_avg_Resale_price_st
plt.show()
```



In [ ]: