# Port Moody Real Estate

STAT 410

Yong Kuk Lee (#301151124)
Heeju Choi (#301341550)

# PURPOSE

Yong Kuk and Heeju are multi-billionaires seeking to invest the entire city of Port Moody. Our interest is to find a reasonable estimation of total value of real estate in Port Moody in order to figure out the total budget needed for us to purchase the entire city of Port Moody.

---

# SURVEY METHODOLOGY

**Target Population**

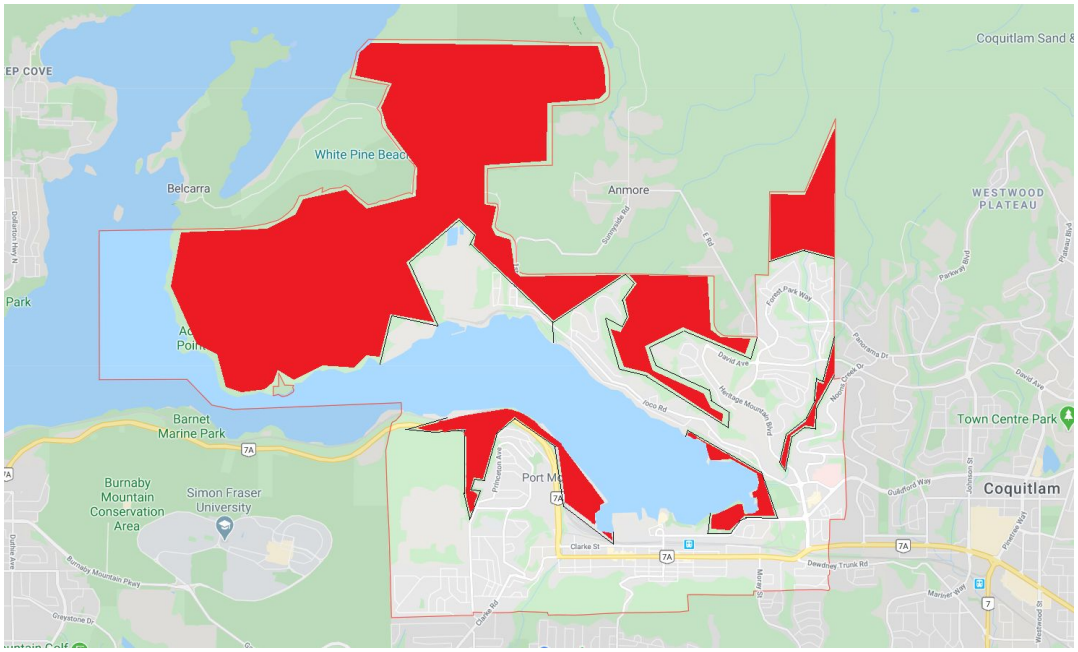
All real estate properties in Port Moody where assessment values have been assigned.

**Population to be Sampled**

All accessible real estate properties information from the BC Assessment website (https://www.bcassessment.ca)

*The target population and the population to be sampled are the same in this survey.

**Sampling Frame**



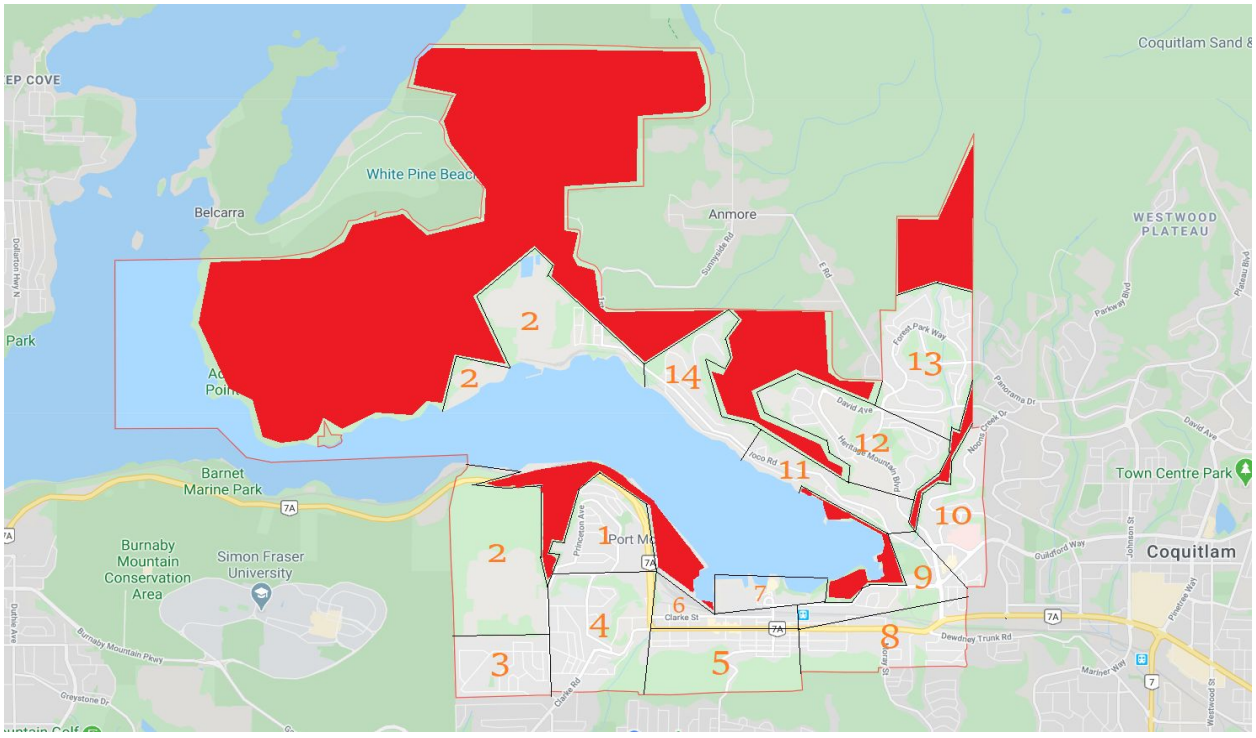

The inside of outer red boundary indicates the land of Port Moody. However, we have ignored the areas coloured in red since they are filled with mountains and properties without assessments where we cannot conduct the sampling. Therefore, areas within the black boundary lines are our sampling frame where there are different densities of the houses.

**Survey Design**

The total value of real estates in Port Moody is estimated using **<u>two-stage cluster sampling</u>.**

Areas in Port Moody is divided into 14 clusters in a criteria of whether the clusters are located near river or not. (ie. Cluster 11 is close to river, Cluster 13 is far from river, and Cluster 12 is located in the middle)

The clusters have been divided as shown below:




We know that the more the samples we have, the better the precision we get. So, we tried to sample as many clusters as possible and we ended up selecting 6 clusters; 6 clusters out of 14 clusters provide an approximately 43 % coverage of the clusters. 6 clusters are randomly selected to conduct the sampling using 'sample(1:14,6)' where it uses SRSWOR; Cluster 2, Cluster 7, Cluster 9, Cluster 10, Cluster 11, and Cluster 13 have been randomly selected among 14 clusters.

Then, we tried to sample as many samples as possible and ended up to randomly select 50 properties for each cluster. Each cluster will run SRS using R to find out which of the assessed properties to be considered as one of our samples; each property will have its own unique number and will be sampled by using "sample()" in R. For instance, we will run "sample(1:M_i,50)" where we have M_i numbers of properties in order to figure out which of 50 properties to be sampled.

Based on the each 50 properties we have sampled, we can estimate the total value and its variance.

# ASSUMPTION

1) There is an equal probability to select a cluster; equally-likely to select a cluster.
2) There is an equal probability to select each sample within a cluster until the number of ssu equals to 50.
3) Our samples are selected by SRSWOR (Simple Random Sampling Without Replacement).
4) Properties that are not registered on the BC Assessment have not taken into account.

# PILOT SURVEY

### Definition of 'Port Moody Real Estate'

Real Estates in Port Moody refer to all properties that possess at least one value from the BC Assessment; land value or building value.

### R software conducting the Pilot Survey

```
##Pilot Study
#Randomly select the clusters to be sampled from the samples we are going to use.
sample(c(2,7,9,10,11,13),3)

#Number of ssus in psu_i where i=2,7,9; M_i.
n2<-c(rep(147,50),rep(2998,50),rep(1217,50))

#Indice of data of each cluster.
i2<- c(51:100)
i9<- c(401:450)
i13<- c(601:650)

#Save indice of sampled data
samps <- c(i2,i7,i9,i10,i11,i13)

#Save the selected clusters to a new dataset called 'prop'
prop <- data[samps,]

#Combine info altogether
prop <- cbind(prop,n1,n2 )

#Two-stage cluster sampling
prop.des <- svydesign(data=prop, id=~id1+id2,fpc=~n1+n2)

#Estimatd total value
est.total=svytotal(~value,design=prop.des)
est.total

#95% confidence interval
confint(est.total)

#Variance
vcov(est.total)
```

```
> est.total
              total          SE
value 6.3283e+10 3.9547e+10
> #95% confidence interval
> confint(est.total)
                2.5 %       97.5 %
value -14228097565 140793601581
> #Variance
> vcov(est.total)
              value
value 1.563971e+21
```

### Findings from the Pilot Survey

The pilot survey has been conducted with 3 randomly selected clusters, Cluster 2, Cluster 9, and Cluster 13, among 14 clusters, then randomly have sampled 50 properties from each clusters. The estimated total value of properties in Port Moody is $ 63,283,000,000 with the standard error of $ 39,547,000,000.
Then, it has 95% confidence interval [$ -14,228,097,565 , $ 140,793,601,581]

It has a variance value of $ 1.563871e+21.

*The variance of the estimated total is variability between clusters and the variability of samples within each cluster.

**Suggestion for Survey**

We have found that the standard error of the estimated total is noticeably large. Another noticeable fact is that the 95% confidence interval contains a unreliable negative value; the estimated total value must be within positive real number. Therefore, our goal in actual survey is to reduce the variance and narrow 95% confidence interval. In order to do so, we would like to use more clusters in the actual survey.

---

# OUR SAMPLED PROPERTY DATASET

By conducting SRSWOR with a sample size of 50, we have collected data of 14 clusters with 50 properties each.

| | id1 | id2 | value |
|---|---|---|---|
| 1 | 1 | 1 | 5291000 |
| 2 | 1 | 2 | 1566000 |
| 3 | 1 | 3 | 152000 |
| 4 | 1 | 4 | 1212000 |
| 5 | 1 | 5 | 1331000 |
| 6 | 1 | 6 | 1442000 |

head(data); first six observations

| | id1 | id2 | value |
|---|---|---|---|
| 695 | 14 | 45 | 981000 |
| 696 | 14 | 46 | 1071000 |
| 697 | 14 | 47 | 1202000 |
| 698 | 14 | 48 | 2016000 |
| 699 | 14 | 49 | 1130000 |
| 700 | 14 | 50 | 1365000 |

tail(data); last six observations

#id1 infers the cluster number (index)
#id2 infers the index of second stage unit
#value infers the total value (land value + building value) for each property

# SUMMARY OF FINDINGS

The estimated total value of all properties in Port Moody is $ 38,110,000,000 with the standard error of $ 18,201,000,000.
Then, it has 95% confidence interval [$ 2,436,886,904 , $ 73,782,157,956]

It has a variance value of $ 3.31264e+20
*The variance of the estimated total is variability between clusters and the variability of each value of sample within each clusters.

**Comparison with the Pilot Survey**

The standard error of the estimated total, its 95% confidence interval, and its variance have been reduced by sampling additional clusters. Thereby, it is reasonable to say that we have more precise estimation compared to the pilot survey.

# DETAILED DESCRIPTION

**Graphical Statistics using Boxplot**

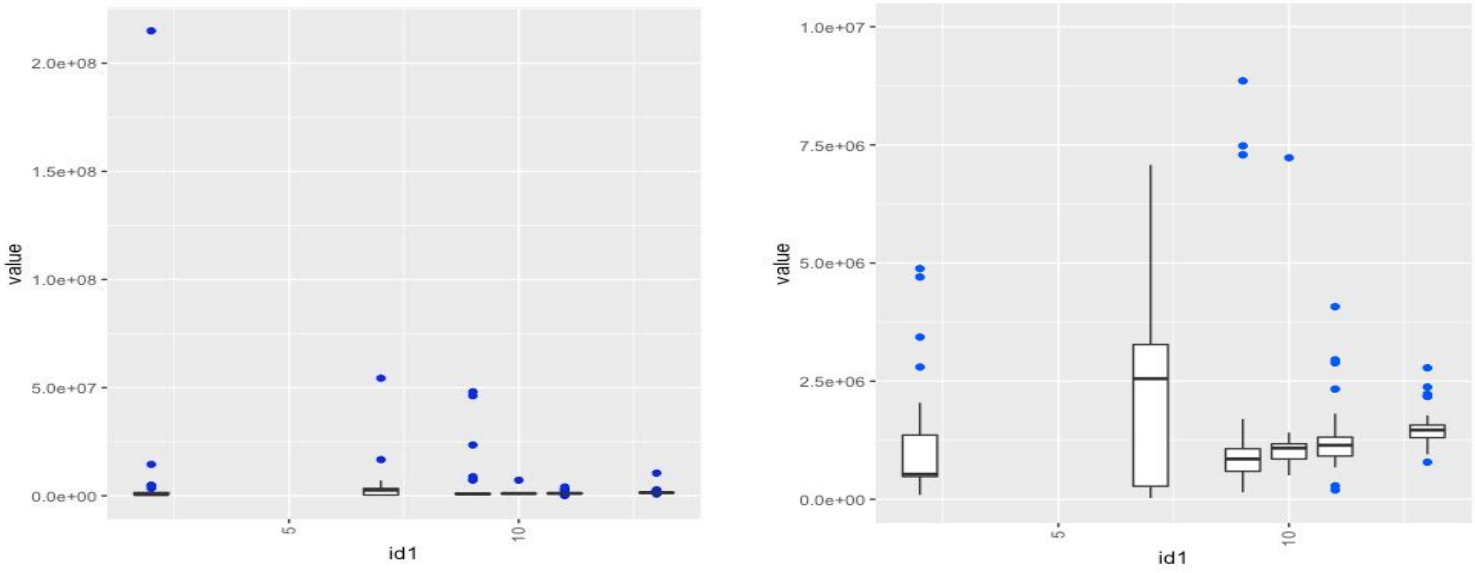



Figure a. (left boxplot)) Boxplot of sampled properties
Figure b. (right boxplot) Boxplot of sampled properties with limited y-range

Two boxplots are indicating the properties' values for the selected clusters (2,7,9,10,11, and 13). They look the same, however, the figure b) contains limited y range so that the other samples can be seen clearly; few huge outliers are ignored.

*Note: The figures indicate that the huge outliers can possibly impact the variance, however we would accept as it is because it certainly reflects the total value of Port Moody. Instead, we have sampled more clusters in order to reduce the variance of the estimation.

**R software conducting the Two-stage Cluster Sampling**

```
#Randomly choose clusters to be sampled
sample(1:14,4)

#Total number of clusters, N
n1<-14

#Number of ssus in psu_i where i = 2,7,9,10,11,13
n2<-c(rep(147,50),rep(65,50),rep(2998,50),rep(1398,50),rep(770,50),rep(1217,50))

#Indice of data of each cluster.
i2<- c(51:100)
i7<- c(301:350)
i9<- c(401:450)
i10 <- c(451:500)
i11<- c(501:550)
i13<- c(601:650)

#Save indice of sampled data
samps <- c(i2,i7,i9,i10,i11,i13)

#Save the selected clusters to a new dataset called 'prop'
prop <- data[samps,]

#Combine info altogether
prop <- cbind(prop,n1,n2 )

#Two-stage cluster sampling
prop.des <- svydesign(data=prop, id=~id1+id2,fpc=~n1+n2)

#Estimatd total value
est.total=svytotal(~value,design=prop.des)
est.total

#95% confidence interval
confint(est.total)

#Variance
vcov(est.total)
```

```
> est.total
          total         SE
value 3.811e+10 1.8201e+10
> #95% confidence level
> confint(est.total)
           2.5 %      97.5 %
value 2436886904 73782157956
> #variance
> vcov(est.total)
             value
value 3.31264e+20
```

step 1. Randomly select 6 clusters to sample
#Cluster 2, Cluster 7, Cluster 9, Cluster 11, and Cluster 13 are randomly selected.

step 2. Sample each selected cluster and gather those data into one place

step 3. Set the total number of clusters in population
#There is 14 clusters in population.

step 4. Set the number of properties within each cluster
#147 properties in Cluster 2, 65 properties in Cluster 7, 2998 properties in Cluster 9, 1398 properties in Cluster 10, 770 properties in Cluster 11, 1217 properties in Cluster 13.

step 5. Bind all needed information such as the selected properties, the total number of clusters, and the total number of properties within each cluster.

step 6. Design two-stage cluster sampling
#'id' indicates the identifier for psu and the second-stage unit respectively.
#'fpc' indicates the number of psu (cluster) in the population and the number of second stage units in each psu respectively.

step 7. Estimate the total value
#Since we would like properties' value to be analyzed in order to get the total value, we have used 'svytotal' function with 'value' variable; as a result, we have finally estimated the total value of Port Moody to be $ 38,110,000,000 with their corresponding standard error and 95% confidence interval.

## PROBLEM

1) There are some samples with outlying values; as we can see from Cluster 2, one of them has an outstanding value that can possibly cause a bias over our inferences about the entire population, causing a higher variability.
2) The total number of properties is not given from the BC Assessment, so we had to count all the properties manually; stacks of mis-counts can influence our model, so we had to be precise on what we were counting.

## CONCLUSIONS & SUGGESTIONS

The estimation of total value of real estates in Port Moody is found out to be $ 38,110,000,000, which is a reasonable amount for us to handle, with a standard error of $ 18,201,000,000 according to our calculation. In other words, we are 95% confident that the true value lies between $ 2,436,886,904 and $ 73,782,157,956. So, as far as we can potentially pay up to $ 73,782,157,956, we will be able to purchase the entire city. We know that the clustering sampling is prone to biases but it requires fewer resources compared to simple random sampling or stratified random sampling; so it is cost-effective, but we would still suggest to use as many clusters as possible to reduce the variability/uncertainty if more manpower is available.