Zoning out: an analysis of zoning and property values in Washington, D.C.

Brian Bontempo, Derrick Lee, and Aaron R. Williams

Sampling Frame

Download the data

```
library(tidyverse)
library(knitr)
```

```
# file path to csv with addresses
aru_file_path <-
  "https://opendata.arcgis.com/datasets/c3c0ae91dca54c5d9ce56962fa0dd645_68.csv"

ap_file_path <-
  "https://opendata.arcgis.com/datasets/aa514416aaf74fdc94748f1e56e7cc8a_0.csv"

# create a directory for downloading the data
if (!dir.exists("data/")) {
  dir.create("data")
}

# if the data doesn't already exist, download the data
if (!file.exists("data/aru.csv")) {
  download.file(aru_file_path, "data/aru.csv")
}

if (!file.exists("data/ap.csv")) {
  download.file(ap_file_path, "data/ap.csv")
}
```

Address Residential Units

The first dataset is Address Residential Units

The dataset does not contain a variable for quadrant, so we extract quadrant from the full address.

```
aru <- read_csv("data/aru.csv") %>%
  rename_all(tolower) %>%
  select(unit_id, address_id, fulladdress, status, unitnum, unittype)

# extract quadrant
aru <- aru %>%
  mutate(quadrant = str_sub(fulladdress, start = -2, end = -1))
```

Address Residential Units contains residential units with status set to "RETIRED". We drop these cases as well.

```
count(aru, status) %>%
  kable()
```

status	n
ACTIVE	244046
ASSIGNED	47
RETIRE	7087

```
aru <- aru %>%
  filter(status != "RETIRE")
```

Address Points

```
# load the data and convert the variable names to lower case
ap <- read_csv("data/ap.csv", guess_max = 10000) %>%
  rename_all(tolower) %>%
  select(address_id, status, type_, entrancetype, quadrant, fulladdress,
         objectid_1, assessment_nbhd, cfsa_name, census_tract, vote_prcnct,
         ward, zipcode, anc, census_block, census_blockgroup, latitude,
         longitude, active_res_unit_count, res_type, active_res_occupancy_count)
```

Address Points contains residential units, non-residential units, and mixed-use units. Residential units and mixed-use units contain residences that belong to our sampling frame. We drop non-residential units.

```
count(ap, res_type) %>%
  kable()
```

res_type	n
MIXED USE	473
NON RESIDENTIAL	15807
RESIDENTIAL	131370

```
ap <- ap %>%
  filter(res_type != "NON RESIDENTIAL")
```

Address points contains residential units with status set to "RETIRED". We drop these cases as well.

```r
count(ap, status) %>%
  kable()
```

status	n
ACTIVE	128490
ASSIGNED	668
RETIRE	2675
TEMPORARY	10

```r
ap <- ap %>%
  filter(status != "RETIRE")
```

After the above filtering, there are 98 observations from Address Points and 3,706 observations in Address Residential Units that have missing addresses. We investigated joining the two datasets on `address_id` to fill in the address but all records missing an address in one dataset were missing an address in the other dataset.

We dropped the missing values which represented about 1.5 percent of observations in Address Residential Units and 0.07 percent of observations in Address Points.

```r
ap <- ap %>%
  filter(!is.na(fulladdress))

aru <- aru %>%
  filter(!is.na(fulladdress))
```

Merge variables

Address Points has interesting variables not present in Address Residential Units. So we merge the Address Points dataset with the Address Residential Units dataset. The join works for all but 572 cases, most of which are in a new building at the Wharf.

```r
aru_expanded <- aru %>%
  select(-status) %>%
  left_join(ap, by = c("fulladdress", "address_id")) %>%
  select(quadrant = quadrant.x, everything(), -quadrant.y)

anti_join(aru, ap, by = c("fulladdress", "address_id"))
```

```
## # A tibble: 572 x 7
##    unit_id address_id fulladdress         status unitnum unittype quadrant
##      <dbl>      <dbl> <chr>               <chr>  <chr>   <chr>    <chr>
##  1  223379     276680 600 WATER STREET SW ACTIVE 6-12    RENTAL   SW
##  2  223380     276680 600 WATER STREET SW ACTIVE 6-13    RENTAL   SW
##  3  223381     276680 600 WATER STREET SW ACTIVE 6-14    RENTAL   SW
##  4  223384     276680 600 WATER STREET SW ACTIVE 1-1     RENTAL   SW
```

```
##  5  223389     276680 600 WATER STREET SW ACTIVE 1-6     RENTAL   SW
##  6  223392     276680 600 WATER STREET SW ACTIVE 1-9     RENTAL   SW
##  7  223494     276680 600 WATER STREET SW ACTIVE 8-16    RENTAL   SW
##  8  223497     276680 600 WATER STREET SW ACTIVE 9-3     RENTAL   SW
##  9  223503     276680 600 WATER STREET SW ACTIVE 9-9     RENTAL   SW
## 10  223508     276680 600 WATER STREET SW ACTIVE 9-14    RENTAL   SW
## # ... with 562 more rows
```

```
rm(aru)
```

Combination

Next, we need to drop addresses in the Address Points dataset that exist in the Address Residential Units dataset so we don't over count addresses in multi-dwelling units.

```
ap <- ap %>%
  filter(!address_id %in% unique(aru_expanded$address_id))
```

Finally, we can combine the two datasets to create a sampling frame that contains approximately every residential address in Washington D.C.

```
sampling_frame <- bind_rows(ap, aru_expanded)

rm(ap, aru_expanded)

#summarize_all(addresses, list(~sum(is.na(.))))

write_csv(sampling_frame, "sampling_frame.csv")
```

Pilot survey

```
set.seed(20190714)

pilot_sample <- sampling_frame %>%
  group_by(quadrant) %>%
  sample_n(25)

write_csv(pilot_sample, "data/pilot_sample.csv")

rm(pilot_sample)
```

```
# load the completed pilot survey and clean the values
pilot_sample <- read_csv("data/pilot_sample_completed.csv") %>%
  mutate(land_value = ifelse(!is.na(rf_land_value),
                             rf_land_value,
                             land_value),
```

```
        improvement_value = ifelse(!is.na(rf_improvement_value),
                                   rf_improvement_value,
                                   improvement_value)) %>%
  mutate(property_value = land_value + improvement_value) %>%
  mutate(property_value = ifelse(unittype == "RENTAL" &
                                   active_res_occupancy_count > 4 &
                                   property_value > 500000,
                                 property_value / active_res_occupancy_count,
                                 property_value
                                 ))
```

```
pilot_sample %>%
  summarize(mean = mean(property_value, na.rm = TRUE),
            s_squared_h = var(property_value, na.rm = TRUE),
            missing_prop = mean(is.na(property_value))) %>%
  kable(caption = "Pilot survey summary statistics")
```

Table 4: Pilot survey summary statistics

mean	s_squared_h	missing_prop
454852.5	259886899569	0.17

```
pilot_sample %>%
  group_by(quadrant) %>%
  summarize(mean = mean(property_value, na.rm = TRUE),
            s_squared_h = var(property_value, na.rm = TRUE),
            missing_prop = mean(is.na(property_value))) %>%
  kable(caption = "Pilot survey summary statistics by quadrant")
```

Table 5: Pilot survey summary statistics by quadrant

quadrant	mean	s_squared_h	missing_prop
NE	408489.5	55231295979	0.08
NW	781327.7	715270634804	0.12
SE	305901.6	71519718277	0.28
SW	283103.1	25025018879	0.20

Picking stratum sizes

Condition 1: Sample mean

We begin with a derivation of Exact Optimal Sample Allocation for $\bar{y}$.

Decomposition of $V(\bar{y}_h)$:

By Wright (12.4), $V(\bar{y}_{str}) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 V(\bar{y}_h) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$

$V(\bar{y}_h) = (\frac{N_h}{N})^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$

$V(\bar{y}_h) = (\frac{N_h^2}{N^2})(1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}$

$V(\bar{y}_h) = (\frac{N_h^2 S_h^2}{N^2})(\frac{1}{n_h}) - \frac{N_h^2 n_h S_h^2}{N^2 N_h n_h}$

$V(\bar{y}_h) = (\frac{N_h^2 S_h^2}{N^2})(\frac{1}{n_h}) - \frac{N_h S_h^2}{N^2}$

$V(\bar{y}_h) = (\frac{N_h^2 S_h^2}{N^2})(1 - \frac{1}{1 \cdot 2} - \frac{1}{2 \cdot 3} - ... - \frac{1}{n_h(n_h - 1)}) - \frac{N_h S_h^2}{N^2}$

$V(\bar{y}_h) = \frac{N_h(N_h - 1) S_h^2}{N^2} - \frac{N_h^2 S_h^2}{N^2 \cdot 1 \cdot 2} - \frac{N_h^2 S_h^2}{N^2 \cdot 2 \cdot 3} - ... - - \frac{N_h^2 S_h^2}{N^2 n_h(n_h - 1)}$

Decomposition of $V(\bar{y}_{str})$

$V(\bar{y}_{str}) = \sum_{h=1}^{H} \frac{N_h(N_h - 1) S_h^2}{N^2}$

$- \frac{N_1^2 S_1^2}{N^2 \cdot 1 \cdot 2} - \frac{N_1^2 S_1^2}{N^2 \cdot 2 \cdot 3} - ... - - \frac{N_1^2 S_1^2}{N^2 n_1(n_1 - 1)}$

...

$- \frac{N_h^2 S_h^2}{N^2 \cdot 1 \cdot 2} - \frac{N_h^2 S_h^2}{N^2 \cdot 2 \cdot 3} - ... - - \frac{N_h^2 S_h^2}{N^2 n_h(n_h - 1)}$

...

$- \frac{N_H^2 S_H^2}{N^2 \cdot 1 \cdot 2} - \frac{N_H^2 S_H^2}{N^2 \cdot 2 \cdot 3} - ... - - \frac{N_H^2 S_H^2}{N^2 n_H(n_H - 1)}$

For a desired bound V_0 on the sampling variance $V(\bar{y}_{str})$, we may find an optimal allocation using the following algorithm:

1) Assign, for each stratum, 1 unit to be selected for the sample.

2) Fill in the following table and number these values starting from 1, in decreasing order.

$\frac{N_1^2 S_1^2}{N^2 \cdot 1 \cdot 2}$	$\frac{N_1^2 S_1^2}{n^2 \cdot 2 \cdot 3}$	$\frac{N_1^2 S_1^2}{N^2 \cdot 3 \cdot 4}$	...
$\frac{N_2^2 S_2^2}{N^2 \cdot 1 \cdot 2}$	$\frac{N_2^2 S_2^2}{N^2 \cdot 2 \cdot 3}$	$\frac{N_2^2 S_2^2}{N^2 \cdot 3 \cdot 4}$	...
.	.	.	...
.	.	.	...
.	.	.	...
$\frac{N_H^2 S_H^2}{N^2 \cdot 1 \cdot 2}$	$\frac{N_H^2 S_H^2}{N^2 \cdot 2 \cdot 3}$	$\frac{N_H^2 S_H^2}{N^2 \cdot 3 \cdot 4}$	...

3) Since the initial allocation is $(n_{11}, n_{21}, ..., n_{H1}) = (1, 1, ..., 1)$, compute $V(\bar{y}_{str} | n_{11} = 1, n_{21} = 1, ..., n_{H1} = 1) = \sum_{h=1}^{H} \frac{N_h(N_h - 1) S_h^2}{N^2}$

4) Pick value (1) from the table and increase the associated stratum's sample

size by 1, o that the updated allocation is $(n_{12}, n_{22}, ..., n_{H2})$, where exactly one of the n_{h2}'s is equal to 2 and the rest are equal to 1. Then, compute $V(\bar{y}_{str}|n_{12}, ..., n_{H2} = V(\bar{y}_{str}|n_{11}, ..., n_{H1}) - \frac{1}{N^2}$ where "(1)" represents the largest value from the table. If $V(\bar{y}_{str}|N_{12}, ..., n_{H2} \leq V_0$, then stop with $n_1 = n_{12}, ..., N_H = N_{H2}$. Otherwise, go to step 5.

5) Pick value (2) from the table and increase the associated stratum's sample size by 1, so that the updated allocation is $(n_{13}, ..., n_{H3})$. Then compute $V(\bar{y}_{str}|n_{13}, ..., n_{H3}) = V(\bar{y}_{str}|n_{12}, ..., n_{H2} - \frac{(2)}{N^2}$, where "(2)" represents the second value from the table. If $V(\bar{y}_{str}|n_{13}, ..., N_H = n_{H3}$. Otherwise, continue until step j, where $V(\bar{y}_s tr|n_{1j}, ..., n_{Hj}) \leq V_0$. The final allocation is $n_{1j}, ..., n_{Hj}$ and $n = n_{1j} + \cdots + n_{Hj}$.

```
# find Nh and s2 for each strata
# (1) and (2)
s_squared_h <- pilot_sample %>%
  group_by(stratum = quadrant) %>%
  summarize(s_squared_h = var(property_value, na.rm = TRUE),
            missing_prop = mean(is.na(property_value)))

Nh <- sampling_frame %>%
  count(stratum = quadrant) %>%
  rename(Nh = n)

strata <- left_join(s_squared_h, Nh, by = "stratum") %>%
  # adjust N because of missingness
  mutate(Nh = Nh * (1 - missing_prop)) %>%
  mutate(N = sum(Nh))

rm(s_squared_h, Nh)

kable(strata)
```

stratum	s_squared_h	missing_prop	Nh	N
NE	55231295979	0.08	68953.08	297153.2
NW	715270634804	0.12	166931.60	297153.2
SE	71519718277	0.28	49423.68	297153.2
SW	25025018879	0.20	11844.80	297153.2

Step 3: $\hat{V}(\bar{y}|1, 1, 1, 1) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{N_h - n_h}{N_H} \frac{s_h^2}{n_h} = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{N_h - 1}{N_H} \frac{s_h^2}{1}$

(Wright 12.5)

```
# Let the initial allocation be (n_11, n_21, n_31, n_41) = (1, 1, 1, 1)
# (3) and (6)
starting_variance <- strata %>%
  mutate(strata_variance = Nh * (Nh - 1) * s_squared_h / N^2)
```

```r
kable(starting_variance)
```

stratum	s_squared_h	missing_prop	Nh	N	strata_variance
NE	55231295979	0.08	68953.08	297153.2	2973894581
NW	715270634804	0.12	166931.60	297153.2	225727358777
SE	71519718277	0.28	49423.68	297153.2	1978456290
SW	25025018879	0.20	11844.80	297153.2	39758730

```r
starting_variance <- starting_variance %>%
  summarize(V = sum(strata_variance)) %>%
  pull()

starting_variance
```

```
## [1] 230719468378
```

Step 3:

Priority value $= \frac{N_1^2 \cdot s_1^2}{N_1^2 \cdot n_h(n_h - 1)}$

```r
# create a table of priority values
# (4) and (5)
n_strata <-
  tibble(stratum = rep(strata$stratum, strata$Nh)) %>%
  group_by(stratum) %>%
  mutate(n = row_number()) %>%
  ungroup() %>%
  left_join(strata, by = "stratum")

# step 2
priority_values <- n_strata %>%
  group_by(stratum) %>%
  # rewritten to avoid integer overflow
  # mutate(priority_value = (Nh ^ 2 * s_squared_h) / (n * lag(n) * N ^ 2)) %>%
  mutate(priority_value = (Nh ^ 2 / n) * (s_squared_h / lag(n)) * (1 / N ^ 2)) %>%
  ungroup() %>%
  arrange(desc(priority_value))

kable(head(select(priority_values, -missing_prop), n = 10))
```

stratum	n	s_squared_h	Nh	N	priority_value
NW	2	715270634804	166931.6	297153.2	112864355500
NW	3	715270634804	166931.6	297153.2	37621451833
NW	4	715270634804	166931.6	297153.2	18810725917
NW	5	715270634804	166931.6	297153.2	11286435550
NW	6	715270634804	166931.6	297153.2	7524290367
NW	7	715270634804	166931.6	297153.2	5374493119

stratum	n	s_squared_h	Nh	N	priority_value
NW	8	715270634804	166931.6	297153.2	4030869839
NW	9	715270634804	166931.6	297153.2	3135120986
NW	10	715270634804	166931.6	297153.2	2508096789
NW	11	715270634804	166931.6	297153.2	2052079191

Step 4:

```
# (7)
priority_values <- priority_values %>%
  mutate(agg_priority_value = cumsum(priority_value)) %>%
  mutate(marginal_variance = starting_variance - agg_priority_value) %>%
  mutate(marginal_sd = sqrt(marginal_variance))

kable(head(select(priority_values, -missing_prop, -N), n = 100), digits = 0)
```

stratum	n	s_squared_h	Nh	priority_value	agg_priority_value	marginal_variance	marginal_sd
NW	2	715270634804	166932	112864355500	112864355500	117855112878	343300
NW	3	715270634804	166932	37621451833	150485807333	80233661045	283255
NW	4	715270634804	166932	18810725917	169296533250	61422935128	247837
NW	5	715270634804	166932	11286435550	180582968800	50136499578	223912
NW	6	715270634804	166932	7524290367	188107259166	42612209212	206427
NW	7	715270634804	166932	5374493119	193481752285	37237716093	192971
NW	8	715270634804	166932	4030869839	197512622125	33206846253	182227
NW	9	715270634804	166932	3135120986	200647743111	30071725267	173412
NW	10	715270634804	166932	2508096789	203155839900	27563628478	166023
NW	11	715270634804	166932	2052079191	205207919091	25511549287	159723
NW	12	715270634804	166932	1710065992	206917985083	23801483295	154277
NE	2	55231295979	68953	1486968855	208404953938	22314514440	149380
NW	13	715270634804	166932	1446978917	209851932855	20867535523	144456
NW	14	715270634804	166932	1240267643	211092200498	19627267880	140097
NW	15	715270634804	166932	1074898624	212167099122	18552369256	136207
SE	2	71519718277	49424	989248161	213156347282	17563121096	132526
NW	16	715270634804	166932	940536296	214096883578	16622584800	128929
NW	17	715270634804	166932	829884967	214926768545	15792699833	125669
NW	18	715270634804	166932	737675526	215664444071	15055024307	122699
NW	19	715270634804	166932	660025471	216324469542	14394998836	119979
NW	20	715270634804	166932	594022924	216918492466	13800975912	117478
NW	21	715270634804	166932	537449312	217455941778	13263526600	115167
NE	3	55231295979	68953	495656285	217951598063	12767870315	112995
NW	22	715270634804	166932	488590284	218440188346	12279280032	110812
NW	23	715270634804	166932	446104172	218886292518	11833175860	108780
NW	24	715270634804	166932	408928824	219295221342	11424247035	106884
NW	25	715270634804	166932	376214518	219671435861	11048032517	105110
NW	26	715270634804	166932	347274940	220018710801	10700757577	103444
SE	3	71519718277	49424	329749387	220348460188	10371008190	101838
NW	27	715270634804	166932	321550870	220670011058	10049457320	100247
NW	28	715270634804	166932	298582951	220968594009	9750874369	98747
NW	29	715270634804	166932	277991023	221246585033	9472883345	97329
NW	30	715270634804	166932	259458289	221506043321	9213425057	95987

stratum	n	s_squared_h	Nh	priority_value	agg_priority_value	marginal_variance	marginal_sd
NE	4	55231295979	68953	247828143	221753871464	8965596914	94687
NW	31	715270634804	166932	242719044	221996590508	8722877870	93396
NW	32	715270634804	166932	227549104	222224139612	8495328766	92170
NW	33	715270634804	166932	213758249	222437897861	8281570517	91003
NW	34	715270634804	166932	201184234	222639082095	8080386283	89891
NW	35	715270634804	166932	189687992	222828770087	7890698290	88830
NW	36	715270634804	166932	179149771	223007919858	7711548520	87815
NW	37	715270634804	166932	169465999	223177385857	7542082521	86845
SE	4	71519718277	49424	164874693	223342260551	7377207827	85891
NW	38	715270634804	166932	160546736	223502807287	7216661091	84951
NW	39	715270634804	166932	152313570	223655120857	7064347521	84050
NE	5	55231295979	68953	148696886	223803817743	6915650635	83160
NW	40	715270634804	166932	144697892	223948515634	6770952744	82286
NW	41	715270634804	166932	137639458	224086155092	6633313286	81445
NW	42	715270634804	166932	131085198	224217240290	6502228088	80636
NW	43	715270634804	166932	124988212	224342228502	6377239876	79858
NW	44	715270634804	166932	119306930	224461535432	6257932946	79107
NW	45	715270634804	166932	114004399	224575539831	6143928546	78383
NW	46	715270634804	166932	109047686	224684587518	6034880860	77684
NW	47	715270634804	166932	104407359	224788994877	5930473501	77010
NW	48	715270634804	166932	100057053	224889051930	5830416448	76357
NE	6	55231295979	68953	99131257	224988183187	5731285191	75705
SE	5	71519718277	49424	98924816	225087108003	5632360375	75049
NW	49	715270634804	166932	95973091	225183081095	5536387283	74407
NW	50	715270634804	166932	92134168	225275215262	5444253116	73785
NW	51	715270634804	166932	88521063	225363736325	5355732052	73183
NW	52	715270634804	166932	85116407	225448852732	5270615646	72599
NW	53	715270634804	166932	81904467	225530757199	5188711179	72033
NW	54	715270634804	166932	78870968	225609628168	5109840210	71483
NW	55	715270634804	166932	76002933	225685631101	5033837277	70950
NW	56	715270634804	166932	73288543	225758919643	4960548735	70431
NE	7	55231295979	68953	70808041	225829727684	4889740694	69927
NW	57	715270634804	166932	70717015	225900444699	4819023679	69419
NW	58	715270634804	166932	68278497	225968723195	4750745182	68926
NW	59	715270634804	166932	65963972	226034687167	4684781211	68445
SE	6	71519718277	49424	65949877	226100637045	4618831333	67962
NW	60	715270634804	166932	63765173	226164402217	4555066161	67491
NW	61	715270634804	166932	61674511	226226076728	4493391650	67033
NW	62	715270634804	166932	59685011	226285761739	4433706639	66586
NW	63	715270634804	166932	57790249	226343551988	4375916390	66151
NW	64	715270634804	166932	55984303	226399536291	4319932087	65726
NW	65	715270634804	166932	54261709	226453798000	4265670377	65312
NE	8	55231295979	68953	53106031	226506904031	4212564347	64904
NW	66	715270634804	166932	52617415	226559521446	4159946932	64498
NW	67	715270634804	166932	51046746	226610568192	4108900186	64101
NW	68	715270634804	166932	49545371	226660113563	4059354815	63713
NW	69	715270634804	166932	48109273	226708222837	4011245541	63334
SE	7	71519718277	49424	47107055	226755329892	3964138486	62961
NW	70	715270634804	166932	46734723	226802064615	3917403763	62589
NW	71	715270634804	166932	45418252	226847482867	3871985511	62225
NW	72	715270634804	166932	44156634	226891639500	3827828878	61869
NW	73	715270634804	166932	42946863	226934586363	3784882015	61521

stratum	n	s_squared_h	Nh	priority_value	agg_priority_value	marginal_variance	marginal_sd
NW	74	715270634804	166932	41786137	226976372500	3743095878	61181
NE	9	55231295979	68953	41304690	227017677190	3701791188	60842
NW	75	715270634804	166932	40671840	227058349030	3661119348	60507
NW	76	715270634804	166932	39601528	227097950558	3621517820	60179
NW	77	715270634804	166932	38572917	227136523475	3582944902	59858
NW	78	715270634804	166932	37583868	227174107343	3545361035	59543
NW	79	715270634804	166932	36632378	227210739721	3508728657	59235
NW	80	715270634804	166932	35716568	227246456289	3473012089	58932
SE	8	71519718277	49424	35330291	227281786581	3437681797	58632
NW	81	715270634804	166932	34834678	227316621258	3402847120	58334
NW	82	715270634804	166932	33985051	227350606310	3368862068	58042
NW	83	715270634804	166932	33166134	227383772444	3335695934	57755
NE	10	55231295979	68953	33043752	227416816196	3302652181	57469
NW	84	715270634804	166932	32376465	227449192661	3270275717	57186
NW	85	715270634804	166932	31614665	227480807326	3238661052	56909

```
rm(n_strata)
```

```
condition1 <- priority_values %>%
  mutate(stratum = factor(stratum)) %>%
  filter(marginal_variance >= ((0.1 * (mean(pilot_sample$property_value, na.rm = TRUE))) ^ 2))

condition1 <- condition1 %>%
  count(stratum, .drop = FALSE)
```

Condition 2: Sample means within strata

We are interested in comparing $\bar{y}_h$ from the four different quadrants.

$n = \frac{N\sigma^2}{(N-1)\frac{e^2}{z^2_{\frac{\alpha}{2}}}+\sigma^2}$

We can use s^2 from our pilot survey as an unbiased estimate for σ^2.

$n = \frac{Ns^2}{(N-1)\frac{e^2}{z^2_{\frac{\alpha}{2}}}+s^2}$

We want \$50,000 precision at a 90% confidence level for the mean of property value in each strata.

```
condition2 <- strata %>%
  mutate(n = (N * s_squared_h) / ((N - 1) * (100000 ^ 2 / qnorm(0.95) ^ 2) + s_squared_h))

condition2 %>%
  kable()
```

stratum	s_squared_h	missing_prop	Nh	N	n
NE	55231295979	0.08	68953.08	297153.2	14.942366

stratum	s_squared_h	missing_prop	Nh	N	n
NW	715270634804	0.12	166931.60	297153.2	193.394282
SE	71519718277	0.28	49423.68	297153.2	19.348776
SW	25025018879	0.20	11844.80	297153.2	6.770496

Condition 3: Sample proportion

We begin with a derivation of Exact Optimal Sample Allocation for $\hat{p}$.

Decomposition of $V(\hat{p}_{str})$

By Wright (12.14), $V(\hat{p}_{str}) = \sum_{h=1}^{H}(\frac{N_h}{N})^2 V(p_h) = \sum_{h=1}^{H}(\frac{N_h}{N})^2 \frac{N_h - n_h}{N_h - 1}\frac{p(1-p)}{n_h}$

$V(\hat{p}_h) = (\frac{N_h}{N})^2 \frac{N_h - n_h}{N_h - 1}\frac{p(1-p)}{n_h}$

$V(\hat{p}_h) = \frac{N_h^2}{N^2}\frac{N_h}{N_h - 1}\frac{p(1-p)}{n_h} - \frac{N_h^2}{N^2}\frac{n_h}{N_h - 1}\frac{p(1-p)}{n_h}$

$V(\hat{p}_h) = \frac{N_h^2}{N^2}\frac{N_h}{N_h - 1}\frac{p(1-p)}{n_h} - \frac{N_h^2}{N^2}\frac{n_h}{N_h - 1}\frac{p(1-p)}{n_h}$

$V(\hat{p}_h) = \frac{N_h^3 p(1-p)}{N^2(N_h - 1)}\frac{1}{n_h} - \frac{N_h^2 p(1-p)}{N^2(N_h - 1)}$

$V(\hat{p}_h) = \frac{N_h^3 p(1-p)}{N^2(N_h - 1)}(1 - \frac{1}{1\cdot 2} - \frac{1}{2\cdot 3} - \cdots - \frac{1}{n_h(n_h - 1)}) - \frac{N_h^2 p(1-p)}{N^2(N_h - 1)}$

$V(\hat{p}_h) = \frac{N_h^3 p(1-p)}{N^2(N_h - 1)} - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot 1\cdot 2} - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot 2\cdot 3} - \cdots - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot n_h(n_h - 1)} - \frac{N_h^2 p(1-p)}{N^2(N_h - 1)}$

$V(\hat{p}_h) = \frac{(N_h^3 - N_h^2)p(1-p)}{N^2(N_h - 1)} - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot 1\cdot 2} - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot 2\cdot 3} - \cdots - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot n_h(n_h - 1)}$

$V(\hat{p}_h) = \frac{N_h^2(N_h - 1)p(1-p)}{N^2(N_h - 1)} - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot 1\cdot 2} - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot 2\cdot 3} - \cdots - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot n_h(n_h - 1)}$

Decomposition of $V(\hat{p}_{str})$

$V(\hat{p}_{str}) = \sum_{h=1}^{H} \frac{N_h^2(N_h - 1)p(1-p)}{N^2(N_h - 1)}$

$-\frac{N_1^3 p(1-p)}{N^2(N_1 - 1)\cdot 1\cdot 2} - \frac{N_1^3 p(1-p)}{N^2(N_1 - 1)\cdot 2\cdot 3} - \ldots - \frac{N_1^3 p(1-p)}{N^2(N_1 - 1)n_h(n_h - 1)}$

$\ldots$

$-\frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot 1\cdot 2} - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)\cdot 2\cdot 3} - \ldots - \frac{N_h^3 p(1-p)}{N^2(N_h - 1)n_h(n_h - 1)}$

$\ldots$

$-\frac{N_H^3 p(1-p)}{N^2(N_H - 1)\cdot 1\cdot 2} - \frac{N_H^3 p(1-p)}{N^2(N_H - 1)\cdot 2\cdot 3} - \ldots - \frac{N_H^3 p(1-p)}{N^2(N_H - 1)n_h(n_h - 1)}$

For a desired bound on V_0 on the sampling variance $V(\hat{p}_{str})$, we may find an optimal allocation using the following algorithm:

1) Assign, for each stratum, 1 unit to be selected for the sample.

2) Fill in the following table and number these values starting from 1, in decreasing order. We assume $p_h = 0.5$ because that is where the variance reaches its global maximum.

$\frac{\frac{1}{4}N_1^3}{N^2(N_1-1)\cdot 1\cdot 2}$	$\frac{\frac{1}{4}N_1^3}{N^2(N_1-1)\cdot 2\cdot 3}$	$\frac{\frac{1}{4}N_1^3}{N^2(N_1-1)\cdot 3\cdot 4}$	$\cdots$
$\frac{\frac{1}{4}N_2^3}{N^2(N_2-1)\cdot 1\cdot 2}$	$\frac{\frac{1}{4}N_2^3}{N^2(N_2-1)\cdot 2\cdot 3}$	$\frac{\frac{1}{4}N_2^3}{N^2(N_2-1)\cdot 3\cdot 4}$	$\cdots$
.	.	.	$\cdots$
.	.	.	$\cdots$
.	.	.	$\cdots$
$\frac{\frac{1}{4}N_H^3}{N^2(N_H-1)\cdot 1\cdot 2}$	$\frac{\frac{1}{4}N_H^3}{N^2(N_H-1)\cdot 2\cdot 3}$	$\frac{\frac{1}{4}N_H^3}{N^2(N_H-1)\cdot 3\cdot 4}$	$\cdots$

3) Since the initial allocation is $(n_{11}, n_{21}, ..., n_{H1}) = (1, 1, ..., 1)$, compute $V(\hat{p}_{str}|n_{11} = 1, n_{21} = 1, ..., n_{H1} = 1) = \frac{1}{N^2}\sum_{h=1}^{H}((N_h^2 - N_h)S_h^2)$

4) Pick value (1) from the table and increase the associated stratum's sample size by 1, o that the updated allocation is $(n_{12}, n_{22}, ..., n_{H2})$, where exactly one of the n_{h2}'s is equal to 2 and the rest are equal to 1. Then, compute $V(\hat{p}_{str}|n_{12}, ..., n_{H2} = V(\hat{p}_{str}|n_{11}, ..., n_{H1}) - \frac{1}{N^2}$ where "(1)" represents the largest value from the table. If $V(\hat{p}_{str}|N_{12}, ..., n_{H2} \leq V_0$, then stop with $n_1 = n_{12}, ..., N_H = N_{H2}$. Otherwise, go to step 5.

5) Pick value (2) from the table and increase the associated stratum's sample size by 1, so that the updated allocation is $(n_{13}, ..., n_{H3})$. Then compute $V(\hat{p}_{str}|n_{13}, ..., n_{H3}) = V(\hat{p}_{str}|n_{12}, ..., n_{H2} - \frac{(2)}{N^2}$, where "(2)" represents the second value from the table. If $V(\hat{p}_{str}|n_{13}, ..., N_H = n_{H3}$. Otherwise, continue until step j, where $V(\hat{p}_{str}|n_{1j}, ..., n_{Hj}) \leq V_0$. The final allocation is $n_{1j}, ..., n_{Hj})$ and $n = n_{1j} + \cdots + n_{Hj}$.

```
#
strata <- sampling_frame %>%
  count(stratum = quadrant) %>%
  rename(Nh = n) %>%
  mutate(N =sum(Nh),
         s_squared_h = 0.5 * (1 - 0.5))

kable(strata)
```

stratum	Nh	N	s_squared_h
NE	74949	348094	0.25
NW	189695	348094	0.25
SE	68644	348094	0.25
SW	14806	348094	0.25

```
# Let the initial allocation be (n_11, n_21, n_31, n_41) = (1, 1, 1, 1)
# (3) and (6)
starting_variance <- strata %>%
  mutate(strata_variance = (Nh ^ 2 * (Nh - 1) * 0.25) / (N ^ 2 * (Nh - 1)))

kable(starting_variance)
```

stratum	Nh	N	s_squared_h	strata_variance
NE	74949	348094	0.25	0.0115899
NW	189695	348094	0.25	0.0742435
SE	68644	348094	0.25	0.0097219
SW	14806	348094	0.25	0.0004523

```
starting_variance <- starting_variance %>%
  summarize(V = sum(strata_variance)) %>%
  pull()

starting_variance
```

```
## [1] 0.09600763
```

```
# create a table of priority values
# (4) and (5)

n_strata <-
  sampling_frame %>%
  count(quadrant)

n_strata <- tibble(stratum = rep(n_strata$quadrant, n_strata$n)) %>%
  group_by(stratum) %>%
  mutate(n = row_number()) %>%
  left_join(strata, by = "stratum")

# step 2
priority_values <- n_strata %>%
  group_by(stratum) %>%
  mutate(priority_value = (0.25 * Nh ^ 3) / (N ^ 2 * (Nh - 1) * n * lag(n))) %>%
  ungroup() %>%
  arrange(desc(priority_value))

kable(head(priority_values, n = 10))
```

stratum	n	Nh	N	s_squared_h	priority_value
NW	2	189695	348094	0.25	0.0371220
NW	3	189695	348094	0.25	0.0123740
NW	4	189695	348094	0.25	0.0061870
NE	2	74949	348094	0.25	0.0057950
SE	2	68644	348094	0.25	0.0048610
NW	5	189695	348094	0.25	0.0037122
NW	6	189695	348094	0.25	0.0024748
NE	3	74949	348094	0.25	0.0019317
NW	7	189695	348094	0.25	0.0017677
SE	3	68644	348094	0.25	0.0016203

```r
# (7)
priority_values <- priority_values %>%
  mutate(agg_priority_value = cumsum(priority_value)) %>%
  mutate(marginal_variance = starting_variance - agg_priority_value) %>%
  mutate(marginal_sd = sqrt(marginal_variance))

kable(head(select(priority_values, -N), n = 100), align = "l")
```

stratum	n	Nh	s_squared_h	priority_value	agg_priority_value	marginal_variance	marginal_sd
NW	2	189695	0.25	0.0371220	0.0371220	0.0588857	0.2426637
NW	3	189695	0.25	0.0123740	0.0494960	0.0465117	0.2156657
NW	4	189695	0.25	0.0061870	0.0556830	0.0403247	0.2008101
NE	2	74949	0.25	0.0057950	0.0614780	0.0345297	0.1858216
SE	2	68644	0.25	0.0048610	0.0663390	0.0296686	0.1722459
NW	5	189695	0.25	0.0037122	0.0700512	0.0259564	0.1611100
NW	6	189695	0.25	0.0024748	0.0725260	0.0234816	0.1532372
NE	3	74949	0.25	0.0019317	0.0744577	0.0215500	0.1467991
NW	7	189695	0.25	0.0017677	0.0762254	0.0197823	0.1406494
SE	3	68644	0.25	0.0016203	0.0778457	0.0181619	0.1347661
NW	8	189695	0.25	0.0013258	0.0791715	0.0168361	0.1297541
NW	9	189695	0.25	0.0010312	0.0802027	0.0158050	0.1257178
NE	4	74949	0.25	0.0009658	0.0811685	0.0148391	0.1218160
NW	10	189695	0.25	0.0008249	0.0819934	0.0140142	0.1183816
SE	4	68644	0.25	0.0008102	0.0828036	0.0132040	0.1149088
NW	11	189695	0.25	0.0006749	0.0834786	0.0125291	0.1119334
NE	5	74949	0.25	0.0005795	0.0840581	0.0119496	0.1093141
NW	12	189695	0.25	0.0005625	0.0846205	0.0113871	0.1067105
SE	5	68644	0.25	0.0004861	0.0851066	0.0109010	0.1044080
NW	13	189695	0.25	0.0004759	0.0855825	0.0104251	0.1021034
NW	14	189695	0.25	0.0004079	0.0859905	0.0100172	0.1000858
NE	6	74949	0.25	0.0003863	0.0863768	0.0096308	0.0981368
NW	15	189695	0.25	0.0003535	0.0867303	0.0092773	0.0963187
SE	6	68644	0.25	0.0003241	0.0870544	0.0089532	0.0946214
NW	16	189695	0.25	0.0003093	0.0873638	0.0086439	0.0929724
NE	7	74949	0.25	0.0002760	0.0876397	0.0083679	0.0914763
NW	17	189695	0.25	0.0002730	0.0879127	0.0080950	0.0899720
NW	18	189695	0.25	0.0002426	0.0881553	0.0078523	0.0886134
SE	7	68644	0.25	0.0002315	0.0883868	0.0076209	0.0872975
SW	2	14806	0.25	0.0002262	0.0886129	0.0073947	0.0859924
NW	19	189695	0.25	0.0002171	0.0888300	0.0071776	0.0847207
NE	8	74949	0.25	0.0002070	0.0890370	0.0069706	0.0834904
NW	20	189695	0.25	0.0001954	0.0892324	0.0067753	0.0823120
NW	21	189695	0.25	0.0001768	0.0894091	0.0065985	0.0812311
SE	8	68644	0.25	0.0001736	0.0895828	0.0064249	0.0801554
NE	9	74949	0.25	0.0001610	0.0897437	0.0062639	0.0791449
NW	22	189695	0.25	0.0001607	0.0899044	0.0061032	0.0781230
NW	23	189695	0.25	0.0001467	0.0900512	0.0059565	0.0771782
SE	9	68644	0.25	0.0001350	0.0901862	0.0058215	0.0762984
NW	24	189695	0.25	0.0001345	0.0903207	0.0056870	0.0754119
NE	10	74949	0.25	0.0001288	0.0904495	0.0055582	0.0745532
NW	25	189695	0.25	0.0001237	0.0905732	0.0054344	0.0737186
NW	26	189695	0.25	0.0001142	0.0906874	0.0053202	0.0729398

stratum	n	Nh	s_squared_h	priority_value	agg_priority_value	marginal_variance	marginal_sd
SE	10	68644	0.25	0.0001080	0.0907954	0.0052122	0.0721955
NW	27	189695	0.25	0.0001058	0.0909012	0.0051064	0.0714593
NE	11	74949	0.25	0.0001054	0.0910066	0.0050011	0.0707182
NW	28	189695	0.25	0.0000982	0.0911048	0.0049029	0.0700204
NW	29	189695	0.25	0.0000914	0.0911962	0.0048114	0.0693644
SE	11	68644	0.25	0.0000884	0.0912846	0.0047230	0.0687244
NE	12	74949	0.25	0.0000878	0.0913724	0.0046352	0.0680826
NW	30	189695	0.25	0.0000853	0.0914577	0.0045499	0.0674530
NW	31	189695	0.25	0.0000798	0.0915376	0.0044701	0.0668586
SW	3	14806	0.25	0.0000754	0.0916130	0.0043947	0.0662924
NW	32	189695	0.25	0.0000748	0.0916878	0.0043198	0.0657255
NE	13	74949	0.25	0.0000743	0.0917621	0.0042455	0.0651578
SE	12	68644	0.25	0.0000737	0.0918357	0.0041719	0.0645902
NW	33	189695	0.25	0.0000703	0.0919060	0.0041016	0.0640436
NW	34	189695	0.25	0.0000662	0.0919722	0.0040354	0.0635249
NE	14	74949	0.25	0.0000637	0.0920359	0.0039717	0.0630217
NW	35	189695	0.25	0.0000624	0.0920983	0.0039093	0.0625247
SE	13	68644	0.25	0.0000623	0.0921606	0.0038470	0.0620244
NW	36	189695	0.25	0.0000589	0.0922195	0.0037881	0.0615475
NW	37	189695	0.25	0.0000557	0.0922753	0.0037324	0.0610930
NE	15	74949	0.25	0.0000552	0.0923305	0.0036772	0.0606397
SE	14	68644	0.25	0.0000534	0.0923839	0.0036238	0.0601976
NW	38	189695	0.25	0.0000528	0.0924367	0.0035709	0.0597574
NW	39	189695	0.25	0.0000501	0.0924868	0.0035208	0.0593367
NE	16	74949	0.25	0.0000483	0.0925351	0.0034726	0.0589284
NW	40	189695	0.25	0.0000476	0.0925827	0.0034250	0.0585232
SE	15	68644	0.25	0.0000463	0.0926290	0.0033787	0.0581263
NW	41	189695	0.25	0.0000453	0.0926742	0.0033334	0.0577356
NW	42	189695	0.25	0.0000431	0.0927173	0.0032903	0.0573610
NE	17	74949	0.25	0.0000426	0.0927600	0.0032477	0.0569884
NW	43	189695	0.25	0.0000411	0.0928011	0.0032066	0.0566265
SE	16	68644	0.25	0.0000405	0.0928416	0.0031661	0.0562677
NW	44	189695	0.25	0.0000392	0.0928808	0.0031268	0.0559179
NE	18	74949	0.25	0.0000379	0.0929187	0.0030889	0.0555782
SW	4	14806	0.25	0.0000377	0.0929564	0.0030512	0.0552381
NW	45	189695	0.25	0.0000375	0.0929939	0.0030137	0.0548976
NW	46	189695	0.25	0.0000359	0.0930298	0.0029779	0.0545700
SE	17	68644	0.25	0.0000357	0.0930655	0.0029421	0.0542415
NW	47	189695	0.25	0.0000343	0.0930998	0.0029078	0.0539240
NE	19	74949	0.25	0.0000339	0.0931337	0.0028739	0.0536089
NW	48	189695	0.25	0.0000329	0.0931666	0.0028410	0.0533010
SE	18	68644	0.25	0.0000318	0.0931984	0.0028092	0.0530022
NW	49	189695	0.25	0.0000316	0.0932300	0.0027777	0.0527035
NE	20	74949	0.25	0.0000305	0.0932605	0.0027472	0.0524134
NW	50	189695	0.25	0.0000303	0.0932908	0.0027169	0.0521235
NW	51	189695	0.25	0.0000291	0.0933199	0.0026877	0.0518434
SE	19	68644	0.25	0.0000284	0.0933483	0.0026593	0.0515686
NW	52	189695	0.25	0.0000280	0.0933763	0.0026313	0.0512964
NE	21	74949	0.25	0.0000276	0.0934039	0.0026037	0.0510267
NW	53	189695	0.25	0.0000269	0.0934308	0.0025768	0.0507621
NW	54	189695	0.25	0.0000259	0.0934568	0.0025508	0.0505059
SE	20	68644	0.25	0.0000256	0.0934824	0.0025253	0.0502520

stratum	n	Nh	s_squared_h	priority_value	agg_priority_value	marginal_variance	marginal_sd
NE	22	74949	0.25	0.0000251	0.0935075	0.0025002	0.0500017
NW	55	189695	0.25	0.0000250	0.0935325	0.0024752	0.0497511
NW	56	189695	0.25	0.0000241	0.0935566	0.0024511	0.0495083
NW	57	189695	0.25	0.0000233	0.0935798	0.0024278	0.0492728
SE	21	68644	0.25	0.0000231	0.0936030	0.0024047	0.0490374

```
rm(n_strata)
```

```
condition3 <- priority_values %>%
  filter(marginal_variance >= ((0.1 * 0.5) ^ 2))

condition3 <- count(condition3, stratum)
```

Condition 4: Sample proportion within strata

We are interested in comparing $\hat{p}_h$ from the four different quadrants.

$n = \frac{Np(1-p)}{(N-1)\frac{e^2}{z^2_{\frac{\alpha}{2}}}+p(1-p)}$

We can assume that $p = 0.5$.

$n = \frac{\frac{1}{4}N}{(N-1)\frac{e^2}{z^2_{\frac{\alpha}{2}}}+\frac{1}{4}}$

We want 0.1 precision at a 90% confidence level for the mean of proportion with multi-family zoning in each strata.

```
condition4 <- strata %>%
  mutate(n = (N * 0.25) / ((N - 1) * (0.1 ^ 2 / qnorm(0.95) ^ 2) + 0.25))

condition4 %>%
  kable()
```

stratum	Nh	N	s_squared_h	n
NE	74949	348094	0.25	67.62564
NW	189695	348094	0.25	67.62564
SE	68644	348094	0.25	67.62564
SW	14806	348094	0.25	67.62564

Combining the above conditions

We want to sample at a rate that meets the four different requirements from above

1. $V_0 > V(\bar{y}_{str})$ for the sample mean

2. \$50,000 precision at a 90% confidence level for $\bar{y}_h$ in each strata
3. $V_0 > V(\hat{p}_h)$ for the sample proportion
4. 0.1 precision at a 90% confidence level for $\hat{p}$ in each strata

```
tibble(quadrant = condition1$stratum,
       `1.` = condition1$n,
       `2.` = condition2$n,
       `3.` = condition3$n,
       `4.` = condition4$n) %>%
  kable(caption = "Recommended strata sizes across the four conditions")
```

Table 18: Recommended strata sizes across the four conditions

quadrant	1.	2.	3.	4.
NE	14	14.942366	21	67.62564
NW	132	193.394282	53	67.62564
SE	11	19.348776	19	67.62564
SW	1	6.770496	3	67.62564

```
tibble(quadrant = condition1$stratum,
       `1.` = condition1$n,
       `2.` = condition2$n,
       `3.` = condition3$n,
       `4.` = condition4$n) %>%
  gather(key = "key", value = "nh", -quadrant) %>%
  group_by(quadrant) %>%
  summarize(nh = ceiling(max(nh))) %>%
  kable(caption = "Maximum recommended strata sizes across the four conditions")
```

Table 19: Maximum recommended strata sizes across the four conditions

quadrant	nh
NE	68
NW	194
SE	68
SW	68