[Weekly Report]

- Report No.3
- Date: Nov. 21, 2023~ Nov. 28, 2023
- Team: Yang Junseob (2019034639), Ryu Seung Gwon (2019087147)

		Highlights	Self Evaluation
Last Week	Summary	• Reconduct data cleansing and manipulation each file. • Determine the reference area(s) to be compared in ACS	Medium
Current Week	Baseline Goals (Given)	• joining file (geo1, geo2, geo3) and creating integrated DB for analysis • create an integrated database to balanced panel	Medium
	Additional Goals (O.Y.O)	• To create balanced panel data that is consist of three distinct observations	Medium
	Key Issues to Be Resolved	• we need to be eliminating unnecessary duplicated observations in each file. • We must determine a relevant joining strategy such as inner join, left join, right join and full join. • We should closely examine the structure of data; the number of observations includes 207,225.	Medium
	Strategies	• Step 1: cleanse duplicated data in a comprehensive manner. • Step 2: we would do inner join between geo1 and geo2. • Step 3: After, we would do inner join again between geo_temp and geo3. • Step 4: in terms of "PROPERTY_TYPE_YR", we would manipulate unbalanced panel to balanced panel data	Medium
	Results	• After first cleansing, geo1 has 77101obs., geo2 has 192138 obs. and geo3 has 640068obs. • Conducting an inner join between geo1 and geo2, geo_temp has 69768obs. of 5 variables. • Conducting an inner join between geo3 and geo_temp, geo_master has 225032obs. of 11 variables. • Grouping by single pin and remaining pin that has value of 2017, 2018, 2019 each. So, we finally have 207225 obs. in geo_final, type of year has 69065 obs. each. • Finally, we add a new variable LONGITUDE_NBR, LATITUDE_NBR where the value is a random sample of one longitude from the existing LONGITUDE_NBR, LATITUDE_NBR values	Medium
	Implications	• It helps us prepare the data for further analysis or modeling.	Medium
Next Week	Things to Do	• Based upon those queries, we will extract data from the ACS. • Select Model for Analysis	Medium
Remarks	Core Libraries & Packages	• Data Cleansing: dplyr • Grouping: dplyr • Visualization: ggplot2	Medium
	Additional Remarks		

- Remarks
- Work Progress

	Week 1	Week 2	Week 3	Week 4	Week 5
Yang Junseob	EDA	Reconduct EDA	Joining file	-	-
Ryu Seung Gwon	EDA	Reconduct EDA	Joining file	-	-
	Week 6	**Week 7**	**Week 8**	**Week 9**	**Week 10**
Yang Junseob	-	-	-	-	Final Report
Ryu Seung Gwon	-	-	-	-	Final Report

Appendix 1. Results of Joining strategies

Obs.		Inner join	Inner join
geo1	77,101	-	-
geo2	192,138	-	-
geo_temp	-	69,768	-
geo3	640,068	640,068	-
geo_master			225,032

-> We decided that it was difficult to fill the missing values, so we conducted inner join.

Appendix 2. Make every single pin should consist of 3 distinct observations in terms of "PROPERTY_TYPE_YR"

Year	geo_master	geo_final
2017	74,700	69,075
2018	74,935	69,075
2019	75,397	69,075
Total	**225,032**	**207,225**

the goal is to create a balanced panel dataset with unique observations for each PIN3 for the years 2017, 2018, and 2019.

We use group by(PIN3) and filter to retain only those observations where all three years (2017, 2018, 2019) are present for each unique PIN3.

The resulting dataset is named **geo_final**.

Finally, we extracted 69075 obs. each, with a total of **207225 obs**.

Appendix 3. Integrated table of final DB

	PIN3	ASSESSOR_FINAL_ADDR_LN	ASSESSOR_CITY_NM	RESIDENTIAL_IND	PROPERTY_TYPE_YR	LONGITUDE_NBR	LATITUDE_NBR	COMMUNITY_NAME	COMMUNITY_ID	HOUSE_UNIT	PROPERTY_TYPE
1	9.364191e+12	6453 N NORTHWEST HWY	CHICAGO	Y	2017	-87.69490	41.97000	Lincoln Square	4	1	2
2	9.364191e+12	6453 N NORTHWEST HWY	CHICAGO	Y	2018	-87.69490	41.97000	Lincoln Square	4	1	2
3	9.364191e+12	6453 N NORTHWEST HWY	CHICAGO	Y	2019	-87.69490	41.97000	Lincoln Square	4	1	2
4	9.364191e+12	6441 N NORTHWEST HWY	CHICAGO	Y	2017	-87.70000	41.97930	Lincoln Square	4	1	2
5	9.364191e+12	6441 N NORTHWEST HWY	CHICAGO	Y	2018	-87.70000	41.97930	Lincoln Square	4	1	2
6	9.364191e+12	6441 N NORTHWEST HWY	CHICAGO	Y	2019	-87.70000	41.97930	Lincoln Square	4	1	2
7	9.364251e+12	6490 N NORTHWEST HWY	CHICAGO	Y	2017	-87.70210	41.99480	West Ridge	2	1	2
8	9.364251e+12	6490 N NORTHWEST HWY	CHICAGO	Y	2018	-87.70210	41.99480	West Ridge	2	1	2
9	9.364251e+12	6490 N NORTHWEST HWY	CHICAGO	Y	2019	-87.70210	41.99480	West Ridge	2	1	2
10	9.364251e+12	6460 N NORTHWEST HWY	CHICAGO	Y	2017	-87.70810	41.99140	West Ridge	2	1	2
11	9.364251e+12	6460 N NORTHWEST HWY	CHICAGO	Y	2018	-87.70810	41.99140	West Ridge	2	1	2
12	9.364251e+12	6460 N NORTHWEST HWY	CHICAGO	Y	2019	-87.70810	41.99140	West Ridge	2	1	2
13	1.025300e+13	3129 W HOWARD ST	CHICAGO	Y	2017	-87.70780	42.01890	West Ridge	2	1	1
14	1.025300e+13	3129 W HOWARD ST	CHICAGO	Y	2018	-87.70780	42.01890	West Ridge	2	1	1
15	1.025300e+13	3129 W HOWARD ST	CHICAGO	Y	2019	-87.70780	42.01890	West Ridge	2	1	1
16	1.025300e+13	3127 W HOWARD ST	CHICAGO	Y	2017	-87.70780	42.01890	West Ridge	2	1	1
17	1.025300e+13	3127 W HOWARD ST	CHICAGO	Y	2018	-87.70780	42.01890	West Ridge	2	1	1
18	1.025300e+13	3127 W HOWARD ST	CHICAGO	Y	2019	-87.70780	42.01890	West Ridge	2	1	1
19	1.025300e+13	3111 W HOWARD ST	CHICAGO	Y	2017	-87.70720	42.01890	West Ridge	2	1	1
20	1.025300e+13	3111 W HOWARD ST	CHICAGO	Y	2018	-87.70720	42.01890	West Ridge	2	1	1
21	1.025300e+13	3111 W HOWARD ST	CHICAGO	Y	2019	-87.70720	42.01890	West Ridge	2	1	1
22	1.025300e+13	3105 W HOWARD ST	CHICAGO	Y	2017	-87.70690	42.01890	West Ridge	2	1	1

Showing 1 to 23 of 207,225 entries, 11 total columns

Finally, we adjust variable LONGITUDE_NBR, LATITUDE_NBR where the value is a random sample of one longitude from the existing LONGITUDE_NBR, LATITUDE_NBR values.

-> Table shows 1 to 23 of 207,225 entries, 11 total columns