Joshlynn Hilburn

The state of affordable housing in Chicago
Data Cleaning Log

• Opened raw .CSV file in Microsoft Excel
• Saved as ChicagoRentals.XLSX in workbook format
• Created a copy of the source data sheet and renamed it to ChicagoRentals
• Expanded column names to fit all content
• Froze top row to function as header row

• Used Remove Duplicates feature to check if any duplicates existed across all rows; there were none so each row was unique and no two rows shared the exact same combination of information
• To detect blank cells: I highlighted all the cells, pressed F5 to access the Go To menu, and selected all blanks; all blanks were colored as yellow for easy visual distinction
    > It appears that the location data (X Coordinate, Y Coordinate, Latitude, Longitude, and Location) of the entry for Hope Manor Village at 5900-6100 S. Green/Peoria/Sangamon is blank
    > Will flag this issue for now and investigate later, will decide what to do with that row of data towards the end of cleaning
• Selected all data and created a pivot table in a new sheet which I renamed PivotTable
• Used the pivot table in the following ways to clean the sheet of data systematically

• Added Community Area Name to Row Labels and Count of Community Area Names to Values to clean Community Area Names
    > Pressed F7 to access spellcheck to check for typos within pivot table
    > Many Community Area Names are not dictionary words but are what the neighborhood is named; since the Community Area Names are ordered alphabetically, it was easy to spot if a typo was genuinely a typo by the presence of a Community Area Name spelled similarly next to it

Old Community Area Name	Corrected Community Area Name	Reasoning
East Garfiled Park	East Garfield Park	Typo

    > Used find and replace to correct this typo in the ChicagoRentals sheet and refreshed the pivot table to ensure the error was fixed

• Added Community Area Name and Community Area Number in both the Row Labels to verify that each Community Area Name was associated with a single Community Area Number
    > The pivot table showed that Austin had entries with 22 and 25 as Community Area Numbers which is not possible because each Community Area Name can only have one number
    > Sorted ChicagoRentals sheet by Community Area Name in alphabetical order to see that the entry with Madison Renaissance Apts. in the Austin community had 22 instead of 25 as the Community Area Number
    > This is a typo since 22 belongs to Logan Square based on other sheet information
    > Edited this error directly in the sheet and refreshed pivot table to confirm it was fixed; every Community Area Name now had a unique Community Area Number

• Added Property Type in Row Labels and Count of Property Type in Values to clean Property Types
    > At a glance, there are many duplicate categories and typos which can be adjusted to condense to a smaller number of categories while retaining accurate descriptive Property Type names
    > Initially, there are 26 "unique" row labels

> Gave each row label a new name to reflect its content while preserving accuracy; see mapping chart below for corrections

Old Property Type	Corrected Property Type
65+/Supportive	Supportive
ARO	ARO
Artist Housing	Artist
Artist Live/Work Space	Artist
Disabled/Homeless	Disabled
Inter-generational	Multifamily
Multfamily	Multifamily
Multifamily	Multifamily
Multifamily/Artists	Multifamily
Multigfamily	Multifamily
Mutifamily	Multifamily
People with Disabilities	Disabled
Senior	Senior
Senior HUD 202	Senior
Senior LGBTQ	Senior
Seniors	Senior
SRO	SRO
Supportive	Supportive
Supportive Housing	Supportive
Supportive/HIV/AIDS	Supportive
Supportive/Kinship Families	Supportive
Supportive/Males 18-24yrs.	Supportive
Supportive/Teenage Moms	Supportive
Supportive/Veterans	Supportive
Supportive/Youth/Kinship Families	Supportive
Veterans	Veterans

> After using find and replace functions to change fields on the sheet to the new categories, I refreshed the pivot table and confirmed there were now only 8 unique Property Types

> Property Types were now one of eight: Supportive, ARO, Artist, Disabled, Multifamily, Senior, SRO, and Veterans

• Added Property Name in Row Labels and Count of Property Name in Values to clean Property Names

> First, I sorted the Count of Property Name in descending order within the pivot table to check how many and which properties shared the same Property Name

> There were several properties that shared Property Names (ex. Park Douglas is the Property Name for 17 distinct rental units, Lakefront Phase II is the Property Name for 13 distinct rental units, etc.)

> Ownership by the same Management Company but differences in property location explain the similarities in Property Names

> Because there was no consistent and simple way to fix typos and formatting, as spellcheck would flag for Property Names that are not dictionary words, I manually sorted the list of Property Names in alphabetical order and looked at each individually within the pivot table

> I used spellcheck to look for typos and used replace to fix errors with my best judgment

Old Property Name	Corrected Property Name	Reasoning
2556 Armtiage LLC	2556 Armitage LLC	Typo
65th Infantry Boringueneers Apts.	65th Infantry Borinqueneers Apts.	Typo
El ZÃ³calo	El Zócalo	Language formatting
Indepedence Apts.	Independence Apts.	Typo
Park Doulglas	Park Douglas	Typo
The Westner	The Western	Typo

• Added Address in Row Labels and Count of Address in Values to clean Addresses

> Sorted Count of Address by descending order and saw that 9 addresses were shared between two or more properties

> This is a problem because it means either a duplicate was included or multiple unique properties share the same space or building

> The same building or plot of land may host rental units that are sectioned off into their own rental properties

> Looked through the ChicagoRentals sheet manually to see the case for these 9 specific addresses

> For the addresses I left alone, the other information like Property Name and unit number was different which meant it was a unique rental development

Address	Action Taken
5801 N. Pulaski Road	Left alone
2822 W. Jackson Blvd.	Left alone
3541 W. North Ave.	Left alone
6928 N. Wayne Ave.	Left alone
2014 S. Racine Ave.	Left alone
1129 S. Sacramento Ave.	Deleted row with 42 units because correct unit number is 48 based on information from Yellow Pages (https://www.yellowpages.com/chicago-il/mip/dicksons-estates-apartments-512972831)
400 E. 41st St.	Left alone
2626 W. 63rd St.	Left alone
30 W. Cermak Road	Left alone

> Only one of nine addresses being common was the result of a duplicate, which I deleted based on obtaining outside, accurate information

> I used spellcheck to check for typos within the Addresses and used find and replace to fix errors in the ChicagoRentals sheet; the pivot table was refreshed to ensure changes went through

> I also sorted the Row Labels of Addresses in the pivot table in an alphabetical order to ensure that fixing any typos did not lead to a duplicate address

> Spellcheck said that many Addresses were typos because they were not dictionary words; this was ignored on a case-by-case basis because street names do not have to be dictionary words

Old Address	Corrected Address	Reasoning
1254-56 S. Fairfiled Ave.	1254-56 S. Fairfield Ave.	Typo; all other “Fairfiled” typos changed
2642-44 W. 12 th Pl	2642-44 W. 12th Pl	Typo; space between 12 and th
3208 N. Shefield Ave.	3208 N. Sheffield Ave.	Typo
730 N. Milwuakee Ave.	730 N. Milwaukee Ave.	Typo
927 S. Indepenednece Blvd.	927 S. Independence Blvd.	Typo

• Added Zip Code in Row Labels to clean Zip Codes

> Sorted Zip Codes in ascending order to determine the range and any out of bounds locations

> Since all Zip Codes must be in the Chicago area and thus are within a certain range, any outliers could be investigated to see if there was an error in data entry

> The range is 60601 to 66007; 66007 is clearly out of place as it references a region in Kentucky

> The entry with the Zip Code of 66007 is located in the Near West Side Community Area and should actually be 60607 according to rental properties in the same neighborhood; corrected in sheet

• To clean Phone Numbers, I added Phone Number in the Row Labels to quickly observe the variety of numbers

> Decided to make no change to Phone Numbers because there would understandably be duplicates since the same Phone Number is commonly reported for all properties under the same Management Company

> Phone numbers are self-reported and there is no reasonable way to consistently verify that each number is current, working, and accurate

> There are no summary statistics possible to be conducted on phone numbers and their inclusion in the data set is purely for communicative purposes on the consumer end; no analysis is done on phone numbers so cleaning for data analysis purposes was not conducted

• Added Management Company to Row Labels and Count of Management Company to Values to clean Management Companies

> Sorted ChicagoRentals sheet alphabetically by Management Company to match with pivot table view

> Used spellcheck within Management Company column to fix for typos based on the correct English spelling

> Because many Management Companies had names that were not English words, only clear grammatical typos were corrected

Old Management Company	Corrected Management Company	Reasoning
5288 S Blacstone LLC	5288 S Blackstone LLC	Typo; Property Name is "Blackstone"
Campell Street Asset Management Inc.	Campbell Street Asset Management Inc.	Typo
Heartland Houing Inc.	Heartland Housing Inc.	Typo
Heartland Houisng Inc.	Heartland Housing Inc.	Typo
Omni Group	Onni Group	Typo; correct name is "Onni Group" according to company website
Voluinteers of America Illinois	Volunteers of America Illinois	Typo

> Upon closer inspection, I noticed that several Management Companies were the same entity but had different versions of how they were reported or written

> For example, there are 4 rentals having the Management Company "Bickerdike Apartments" and 23 rentals having "Bickerdike Apts."; these appear to be the same company so I decided to combine them based on the dominant formatting

> Corrections were done conservatively for combining the same Management Company only if the difference in the company name was small and the similarity was very clear

> If an equal number of different versions of the same Management Company existed, the more complete and detailed name was chosen as the corrected version

Old Management Company and count	Corrected Management Company and new total count
Bickerdike Apartments (4)	Bickerdike Apts. (27)
Catholic Charities (1)	Catholic Charities Housing Devp. (3)
Deborah's Place (1)	Deborah's Place Ltd. (2)
East Lake Management & Development Corp. and East Lake Management Co. (2 each)	East Lake Management Group, Inc. (10)
Evergreen Real Estate LLC and Evergreen Real Estate Services (1 each)	Evergreen Real Estate Services LLC (4)
Fifield Co. (1)	Fifield Cos. (2)
Flats LLC (1)	FLATS Leasing (4)
Heartland Housing Inc. (2)	Heartland Housing (5)
Leasing and Management Co., Inc. (1)	Leasing & Management Co. Inc (6)
Perlmark Realty Management (1)	Perlmark Realty Management LLC (2)
Related Management (1)	Related Management Co. LP (4)
Senior Lifestyle and Senior Lifestyle Co. (1 each)	Senior Lifestyle Corp. (20)
The Community Builders and The Community Builders, Inc. (2 and 1 respectively)	The Community Builders Inc. (9)
Thresholds (1)	The Thresholds (5)

• Added Units to Row Labels in pivot table to clean Units

> Viewed range of unit numbers to see if the value made sense and to point out possible outliers

> Range of unit numbers is 1 to 534 which is a reasonable value since SRO (Single Room Occupancy) rentals are a category in the dataset and 534 is a standard number for a complex with a large capacity

> No change made to unit numbers

• To clean and verify the accuracy of location data fields (X Coordinate, Y Coordinate, Latitude, Longitude, and Location), I viewed the ranges of each to confirm that they were within locational bounds of the city of Chicago

> Pivot tables with the relevant information added to the Row Labels and sorted alphabetically yielded the range (highest and lowest values)

> For X Coordinate, the range is 1127328.815 to 1201038.059; for Y Coordinate, the range is 1815487.733 to 1949531.171

> For Latitude, the range is 41.64845741 to 42.01714971; for Longitude, the range is -87.80707301 to -87.54012317

> These Latitude and Longitude values match up with the official coordinates of Chicago, which is at 41.8781° N and 87.6298° W

> The units of measurement for the X and Y Coordinates were not described or named within the metadata so the unit used for coordinates is unknown; any analyses of location will be conducted using Latitude and Longitude data

> No changes to location data will be made

• For the rental property with missing location data (Hope Manor Village in Englewood with 36 units at 5900-6100 S. Green/Peoria/Sangamon), this was determined to be the result of the Address referencing three different locations

> For this reason, any analyses of location will exclude this property because there is no singular address to work with

• Since Property Names, Addresses, Phone Numbers, and Locations are shared between two or more properties, any of the existing columns cannot function as a Primary Key since there would be duplicates

• To resolve this, a Composite Key was created in a new column that I named ID that was composed of the Property Name and Address

> Formula used was =D2&" at "&E2 which combined both pieces of information into the format "Property Name at Address"

> Viewing ID in the pivot table showed that 487 unique ID values existed where no two rentals shared the same Property Name and Address combination

> This Primary Key could be used to query information about the dataset

• To verify that each column contained data of the proper data type (Text or Number), a formula was used to check the data type of each column

> Formula used was =TYPE(ChicagoRentals!A$2) in the PivotTable sheet and this was extended across the row to account for all columns

> Used an IF statement under TYPE results with the formula =IF(C5=2,"Text","Number") to determine if the field type was Text or Number

> All expected data types were observed; although Community Area Number and Zip Code are in Number format and better suited as Text, this was left alone because no statistical analyses can be conducted on these two values anyways

• After all cleaning was completed, the column widths were adjusted a final time and the sheet view of ChicagoRentals was exported as ChicagoRentals.csv

• Locked source data sheet and ChicagoRentals (cleaned data) to prevent accidental edits