# Company Specific Housing Data

How to Plan Resourcing for Future Delivery

# GitHub URL

[AnnHayes23/UCDPA\_AnnHayesFinal: Set up new repository from Mac (github.com)](https://github.com/AnnHayes23/UCDPA_AnnHayesFinal)

# Abstract

I used data from my work environment to identify trends in housing types such as apartments versus houses since my company started operating in this sector. I am interested to identify resourcing gaps and or pressure on resourcing based on volume of new homes being delivered in that region.

# Introduction

I thought it would be interesting to work with data that I am familiar with and to see if I could use the learnings from this course to help me with data analysis. Unfortunately, a lot of data that I work with is commercially sensitive and/or personal to tenants so I could only use a limited dataset. To deliver on each milestone, I did use other datasets throughout the assignment. All of the datasets that I used are on my repository [UCDPA\_AnnHayesFinal/data at main · AnnHayes23/UCDPA\_AnnHayesFinal (github.com)](https://github.com/AnnHayes23/UCDPA_AnnHayesFinal/tree/main/data)

# Dataset

I used a csv dataset called **Property Dataset Final** that contains 10,7931 records including a header row and with 13 variables. The data is the property of Clúid Housing which is where I work and Clúid owns or manages these properties. The data types are: int64(2) and object(11).

I requested this data from the data team at Clúid and asked for it to be cleansed of any personal data. I did leave the name of Housing Managers in the dataset (called PropertyAreaOffice) as I wanted to gather some insights into patch sizes for each manager.

This section is included in my code file under the heading **1. Data** and the dataset is also loaded to GitHub for your reference.

**Importing Data**

To demonstrate what I had learned in our class, I imported data from four different locations using the following methods.

1. First, I imported the csv file called, Property Dataset Final, from my hard drive using the code in the box below.

property = pd.read\_csv(r'../data/Property Dataset Final.csv')

1. Then, using Kaggle, I downloaded and imported a dataset called Student Mental health which has 101 rows × 11 columns using the code below.

Student\_MH = pd.read\_csv(r'../data/Student Mental Health.csv')

1. After that I looked through the range of datasets that are available with seaborn, [An introduction to seaborn — seaborn 0.12.2 documentation (pydata.org)](http://seaborn.pydata.org/tutorial/introduction.html#multivariate-views-on-complex-datasets) and I imported a dataset called ‘tips’ which has 244 rows × 7 columns that we had used in class so I was familiar with this data.

import seaborn as sns

tips = sns.load\_dataset("tips")

1. Finally, using an API, I imported a dataset from [All API Collections | RapidAPI](https://rapidapi.com/collections) and converted it to a DataFrame. To do this I had to learn how import requests work, how to get a token and how to use json.loads to covert a json string to a python dictionary. All of the help I needed was available online and I used Stack Overflow to find the answers to the questions I had as they came up. [Convert string to list in python - Stack Overflow](https://stackoverflow.com/questions/14938120/convert-string-to-list-in-python)

import requests

url = "https://unogsng.p.rapidapi.com/genres"

headers = {

"X-RapidAPI-Key": "2f8253dc65mshf5db400c3152ac6p14007djsn78987098a841",

"X-RapidAPI-Host": "unogsng.p.rapidapi.com"

}

response = requests.request("GET", url, headers=headers)

While I first saw the data returning from genres API request, I wasn’t sure what type of data it was so I looked it up.It seemed that it might be json data as it had curly braces used double quotes. To check this assumption I used json.loads and no error was returned so I converted the data to a DataFrame.

Initially I worked in Pycharm but gradually moved to Jupyter because I found it less confusing to see a visual display of the output. However, I did go back to Pycharm when I encountered a problem and tried to figure out exactly which part of the code did not work. Example shown in **Figure 1** is where I could not use a dataset without a token.

Text

Description automatically generated

**Figure 1: Error 403 when I tried to import a dataset without a token.**

This section is included in my code file under the heading **2. Importing Data**.

# Implementation Process

Following the milestone structure for this assignment, I first set out to demonstrate that I understand how to implement the various preparation steps. I included my examples of using NumPy in this first milestone too as I had to use the dataset **Titanic** as **Property Dataset Final** did not have ample numeric values. This section is included in my code file under the heading **3. Preparation of Data**.

Then, I moved on to implement some of the preparation steps on my own dataset to allow me to begin to focus on that data. I first decided to add a column called Province and started to wrestle with how to merge each county under the correct Province. I saw the opportunity to create reusable code and to include conditional formatting and looping for this task. After a lot of errors and research, I came up with the solution below. I created a new file called regions in a python file and in that I was created classes for each county and province. Then I was able to use that file to populate the new column. I felt that sorting by Province would be helpful with some of the analysis.

Text

Description automatically generated

**Figure 2: Reusable code**

Then I dropped columns that I would not be using in the analysis, I renamed other columns to more user friendly names such as Housing Manager, I removed white spaces in the str values using .str.strip() and I filled in blank data using .fillna.

I reduced the dataset even more and decided to only work with homes that were built between 2002 and 2022. This was because the earliest homes were dated back as far as 1930 and the company is only 27 years old. Taking a subset of the most recent twenty years, the data is more robust.

Chart, histogram

Description automatically generated

This section is included in my code file under the heading **4. Analysis of Data**.

# Results

While looking through the data it became apparent that the number of apartments being provided is increasing in the major cities (Cork, Limerick, Galway and Dublin) and that detached and semi-detached houses are far fewer. However, when terraced house are introduced the difference between the delivery of apartments and houses is much less. However if the delivery for 2022 is the beginning of a change, there is a risk of house delivery decreasing in the coming years.

When looking at the data for Housing Managers I realised that a category of Housing Manager existing for Sold/Disposed properties and so I removed that from the dataset.

Text

Description automatically generated

Chart, bar chart

Description automatically generatedThe number of properties being managed by each Housing Manager varies significantly and some most certainly have more capacity than others. Growth targets should consider this and plan to recruit in high volume areas or change strategy and deliver homes in areas where the staff are in place and have capacity to manage additional tenancies. In terms of resourcing, apartments are more intensive on time and costs to maintain.

One other thing that I noticed while doing this exercise is that a data cleansing exercise is needed at my place of work and a data policy is required for standards and rules to be set around how data is captured and archived and/or deleted when no longer required.

With regard to my dataset, I found it very difficult to mirror a lot of the functionality that I learned in the course and now with hindsight I think I would have looked for data with a lot more numeric values so that I could have used more of the pivot table functions and I would have found the visualisation more straightforward too. However, I did enjoy learning about pandas and how quickly data can be sorted, and I am going to continue to use this and hopefully over time become much more proficient and even come back to do the next level of the course.

This section is included in my code file under the heading **5. Data Visualisation**.

# Insights

The insights that I gleaned from this exercise are as follows;

1. The company has experienced growth in delivery of new homes since 2018 and the growth has doubled since 2020.

Chart, histogram

Description automatically generated

1. The supply of detached and semi-detached houses in the major cities of Cork, Galway, Limerick, and Dublin has slowed down and is falling.

Table

Description automatically generated

1. Apartment delivery in the same cities has increased and the company needs to match resourcing should this trend to continue.
2. There is fewer 3+ homes being delivered nationwide each year which could cause an issue for tenants who have large and growing families.

Chart, bar chart

Description automatically generated

1. Housing Managers have different volumes of homes to manage with some reaching capacity and others with capacity to spare. Resourcing will need to be put in place if we continue to deliver homes in the areas where Housing Managers are under pressure due to the volume of homes they manage.

**Machine Learning**

Classification is widely used to analyse data while regression is used on numbers. In the case of my dataset, I believe that classification is the method that would be used for modelling. As an independent variable is a factor that does not change its value even when influenced by other variables, the data in my dataset is independent. Possibly the Housing Manager data could be independent as the person assigned to a region could change if the volume of homes became too much for one person to manage and regions had to be reallocated. It is certain that more advanced analysts could take this data and along with population data could forecast what size homes will be needed in which county, where smaller one bed homes will be needed as the population ages and where more apartments will be built in the future.