**[[1]](#endnote-1)DATS-6103-11**

**Spring 2023**

**Topic Proposal**

**Team 2:** Muhannad Alwhebie, Brian Gulko, Mengfei Hung, Kashyap Nimmagadda

**Topic:** DC Residential Property Sales

Our data comes from Open Data DC and describes the sale history for active properties listed among the District of Columbia’s real property tax assessment roll. The dataset contains about 108,000 rows and 39 columns describing property attributes such as area and number of bedrooms as well as sale information such as sale price and date. Our goal is to use analysis and models to better understand the relationship between these attributes and the effects they have on sale price.

**Attributes of interest:**

1. BATHRM
2. HF\_BATHRM
3. HEAT
4. HEAT\_D
5. NUM\_UNITS
6. ROOMS
7. BEDRM
8. GRADE\_D
9. GRADE
10. GBA
11. LANDAREA
12. SALEDATE
13. PRICE
14. SSL

**SMART Questions:**

1. What are the characteristics of an average residential property?

* Which heating type is the most common in residential properties in this dataset, and what is the percentage of properties with this heating type?
* What is the average number of bathrooms and half-bathrooms in residential properties in this dataset?
* What is the average land area of residential properties in this dataset, and how does this vary by number of bedrooms?
* How has the gross building area of residential properties in this dataset changed over time?

1. Which variables have an impact on sale price, and how strong is that impact?

* Is there a correlation between the number of bedrooms and the sale price of a residential property in this dataset?
* Is there a correlation between the grade and the sale price of a residential property in this dataset?
* Is there a correlation between gross building area and sale price?

1. Did COVID-19 have an impact on residential sale prices? If so how big was that impact?

* How did the average grade of residential properties sold in the District of Columbia change during the COVID-19 pandemic, and was this related to changes in sale price?

We believe these questions can provide valuable insights into the characteristics and trends of residential properties in our dataset.

**Models:** Linear regression1, Multiple linear regression, K-Means clustering2, KNN3

**Dataset Source:** This data comes from Open Data DC’s Computer Assisted Mass Appraisal - Residential dataset which can be found at this link:

<https://opendata.dc.gov/datasets/DCGIS::computer-assisted-mass-appraisal-residential/explore>

Attribute information and descriptions can be found at this link:

<https://www.arcgis.com/sharing/rest/content/items/c5fb3fbe4c694a59a6eef7bf5f8bc49a/info/metadata/metadata.xml?format=default&output=html>

**GitHub Repo:** <https://github.com/kashyapnimmagadda/DATS-6103-TEAM2.git>

1. 1) Linear regression models can be used to understand the relationship between sale price and specific property attributes.

   2) K-Means clustering can help identify groups of properties with similar characteristics.

   3) KNN can be used to predict the sale price of a property based on its attributes and the attributes of nearby properties. [↑](#endnote-ref-1)