Name: Aryan Lunkad
DS210

Final Project Write-up

Project Thesis

The purpose of this project was to analyze the rental property market in Dubai and discover patterns that could help explain pricing trends, market demand, and neighborhood connectivity. I wanted to understand what makes certain neighborhoods stand out and how geographical relationships between properties influence market dynamics. My main goal was to use computational tools like graph theory, predictive modeling, and regression analysis to provide insights that aren't obvious from a simple spreadsheet. For example, I was curious to see whether properties in more connected neighborhoods have higher demand or different pricing trends. At the same time, I wanted to explore how factors like property size and location contribute to rent prices and whether I could predict trends across the dataset.

The research thesis for this project is focused on two key ideas. First, that spatial connectivity plays a major role in determining the desirability of a property, and second that statistical relationships like rent per square foot and property size can reveal pricing benchmarks. By analyzing spatial graphs of properties, I aimed to identify neighborhoods that serve as hubs, mainly areas where high centrality scores suggest strong infrastructure and accessibility. At the same time, predictive modeling and regression analysis allowed me to dig deeper into pricing and demand patterns. Combining these tools gave me a way to bridge the gap between graph theory and real estate economics.

Dataset

Rent	Beds	Baths	Type	Area_in_sqft	Rent_per_sqft	Rent_category	Frequency	Furnishing	Purpose	Posted_date	Age_of_listing_in_days	Location	City	Latitude	Longitude
124000	3	4	Apartment	1785	69.46778711484590	Medium	Yearly	Unfurnished	For Rent	2024-03-07	45	Al Reem Island	Abu Dhabi	24.4935984	54.40784050903400
140000	3	4	Apartment	1422	98.45288326300980	Medium	Yearly	Unfurnished	For Rent	2024-03-08	44	Yas Island	Abu Dhabi	24.4940223	54.6073721
99000	2	3	Apartment	1314	75.34246575342470	Medium	Yearly	Furnished	For Rent	2024-03-21	31	Al Raha Beach	Abu Dhabi	24.48593095	54.60093932996360
220000	3	4	Penthouse	3843	57.24694249284410	High	Yearly	Unfurnished	For Rent	2024-02-24	57	Al Reem Island	Abu Dhabi	24.4935984	54.40784050903400
350000	5	7	Villa	6860	51.02040816326530	High	Yearly	Unfurnished	For Rent	2024-02-16	65	Yas Island	Abu Dhabi	24.4940223	54.6073721
75000	1	1	Apartment	706	106.23229461756400	Medium	Yearly	Furnished	For Rent	2023-12-12	131	Al Reem Island	Abu Dhabi	24.4935984	54.40784050903400
65000	1	1	Apartment	698	93.12320916905440	Low	Yearly	Unfurnished	For Rent	2024-03-11	41	Yas Island	Abu Dhabi	24.4940223	54.6073721
170000	3	4	Townhouse	1989	85.47008547008550	High	Yearly	Unfurnished	For Rent	2024-02-22	59	Yas Island	Abu Dhabi	24.4940223	54.6073721
75000	1	2	Apartment	886	84.65011286681720	Medium	Yearly	Unfurnished	For Rent	2024-04-05	16	Yas Island	Abu Dhabi	24.4940223	54.6073721
160000	2	3	Apartment	1430	111.8881118881120	High	Yearly	Unfurnished	For Rent	2024-04-05	16	Al Reem Island	Abu Dhabi	24.4935984	54.40784050903400
110000	2	2	Apartment	1356	81.12094395280240	Medium	Yearly	Unfurnished	For Rent	2023-01-29	448	Al Bateen	Abu Dhabi	24.2048941	55.6196612
174000	3	4	Townhouse	1816	95.81497797356830	High	Yearly	Unfurnished	For Rent	2024-03-25	27	Yas Island	Abu Dhabi	24.4940223	54.6073721
330000	5	7	Villa	5907	55.865921787709500	High	Yearly	Unfurnished	For Rent	2024-03-22	30	Yas Island	Abu Dhabi	24.4940223	54.6073721
149000	5	6	Villa	3750	39.73333333333330	High	Yearly	Unfurnished	For Rent	2024-04-02	19	Al Reef	Abu Dhabi	24.4664366	54.65689142777210
150000	1	2	Apartment	1022	146.7710371819960	High	Yearly	Furnished	For Rent	2024-03-18	34	The Marina	Abu Dhabi	24.479853	54.319951
80000	2	2	Apartment	1477	54.16384563304000	Medium	Yearly	Unfurnished	For Rent	2024-03-14	38	Al Reef	Abu Dhabi	24.4664366	54.65689142777210
125000	3	4	Apartment	1744	71.6743119266055	Medium	Yearly	Unfurnished	For Rent	2023-12-01	142	Al Reem Island	Abu Dhabi	24.4935984	54.40784050903400
129995	4	5	Villa	2828	45.96711456859970	Medium	Yearly	Unfurnished	For Rent	2024-03-11	41	Al Reef	Abu Dhabi	24.4664366	54.65689142777210
62900	2	2	Apartment	1237	50.84882780921590	Low	Yearly	Unfurnished	For Rent	2024-04-02	19	Al Reef	Abu Dhabi	24.4664366	54.65689142777210
145000	2	4	Townhouse	1821	79.62657880285560	High	Yearly	Unfurnished	For Rent	2024-02-27	54	Yas Island	Abu Dhabi	24.4940223	54.6073721

I chose this dataset because it has the perfect combination of numerical attributes to tackle my research goals. The dataset includes over 73000 rental property listings which is a large enough sample size to ensure strong analysis and meaningful results. What really makes this dataset stand out is the presence of latitude and longitude coordinates for each property. Having this spatial information allowed me to represent the dataset as a graph, where nodes are properties and edges are the geographical connections between them. At the same time, numerical fields like "Rent", "Area in sqft", and "Rent per sqft" provided the basis for statistical models. Another reason this dataset was the right choice is its relevance and real-world applicability. Dubai is a dynamic and fast-growing city where rental markets are influenced by rapid urban development. The dataset captures listings across neighborhoods like "Al Reem Island", "The Marina", and "Yas Island" which are known for their high demand and modern infrastructure. This variety made it ideal for understanding how different areas compare in terms of connectivity and pricing. It also meant that my analysis could reflect real-life trends whether it's identifying hubs of activity or highlighting opportunities in less-saturated areas. By choosing a dataset that's so geographically and economically detailed, I was able to analyze the thesis of this project.

Lastly this dataset stands out because it is clean, well-structured, and easy to work with. The fields are well-defined and the dataset includes essential attributes without unnecessary complexity. For instance, the "Rent" column directly provides annual rental prices. Metrics like these added depth to my analysis and made it possible to connect spatial insights with market behavior.

Code Explanation

The goal of my project was to analyze rental property data in Dubai to identify trends, key neighborhoods, and pricing patterns using graph algorithms, predictive modeling, and regression analysis. I liked this dataset because it combines both spatial data (latitude/longitude) and numerical metrics like rent price, area, and listing age. That combination gave me a good opportunity to apply both graph theory and statistical analysis with results that could actually be meaningful for real estate.

The project starts with data processing. I needed to clean the dataset and extract the most relevant fields. Columns like "Latitude" and "Longitude" were important in constructing the spatial graph, while "Rent" and "Area in sqft" were used for predictive modeling and regression analysis. I also calculated rent per square foot as a derived metric because it's such a common way to compare property prices. During this step I filtered out any properties with missing or invalid coordinates because those would break the graph logic. By the end of this stage, I had a clean and usable dataset with over 73,000 rows.

Next came the graph construction. Here I treated each property as a node in the graph and connected nodes if the properties were within 10 kilometers of each other. To calculate this distance, I implemented the haversine formula which gives accurate great-circle distances between two latitude/longitude points. The result was a graph with over 73,000 nodes and about 350 million edges. Many properties are close together, especially in dense urban areas like "Al Reem Island" or "Yas Island", which I noticed while scanning through the data. This graph became the foundation for the rest of the project because it modeled the relationships between properties in a way that simple spreadsheets or charts couldn't.

With the graph ready, I moved on to centrality analysis. Centrality is a graph measure that shows how important a node is based on its position in the graph. I specifically focused on closeness centrality which tells us how close a property is to all other properties in the graph. I calculated this using a Breadth-First Search (BFS) approach for efficiency. I measured the total distance from each node to all reachable nodes and then inverted that value to get a centrality score. A higher score means the property is well connected and likely in a prime neighborhood. The properties with the highest centrality scores were often in highly connected urban hubs, which makes sense because these areas naturally have more infrastructure, amenities, and accessibility. For example, nodes representing properties in neighborhoods like "The Marina" or "Al Raha Beach" might show higher centrality because of their dense and interconnected layouts.

The next step was predictive modeling. I built a simple model that predicts relative demand or market saturation for properties based on their size and rent price. I didn’t want to overcomplicate this part because the project’s real focus was on graph analysis, but I still wanted something practical. The model works by slightly adjusting the rent per square foot metric to highlight potential hotspots or underpriced properties. For example, if a property in "Khalidiyah" had a rent of 120,000 AED and a smaller area, the adjusted score would help flag whether it aligns with market expectations.

I also implemented a linear regression model to quantify the relationship between "Area in sqft" and "Rent". This step added a nice statistical layer to the project. The regression showed how rent increases with property size, and the results were easy to interpret. For instance, I found that rent increases by a specific amount per additional square foot, which is something landlords could use to benchmark their pricing strategies. The R^2 score I calculated showed that the model fits the data well.

To tie everything together, I used visualizations to make the results more intuitive. I created a centrality map that highlights the most connected properties. The properties with high centrality stood out like natural hubs. I also plotted the regression line for rent versus property sizes which showed a nice upward trend. The visual outputs made the analysis easy to understand for someone who might not be familiar with graphs or centrality scores.

Throughout the project I made sure to test my code rigorously. I wrote test cases to validate the centrality calculations and BFS logic. For example, I tested the code on a small fully connected graph to confirm that all nodes had equal centrality, and I also ran it on a disconnected graph to ensure the BFS handled isolated nodes correctly. These tests were crucial because the dataset was so large that I needed to be confident in the results.

By combining graph algorithms, predictive modeling, and regression, I was able to get a complete picture of Dubai’s rental market. The centrality analysis pinpointed key neighborhoods, the predictive model highlighted potential market trends, and the regression quantified the relationship between size and rent.

Output Summary

```
    Finished `dev` profile [unoptimized + debuginfo] target(s) in 1.23s
     Running `target/debug/DS210_project`
Loading and processing dataset...
Time to process dataset: 170.092917ms
Dataset Loaded: 73742 properties.
Constructing spatial graph...
Graph constructed: 73023 nodes and 350666985 edges.
Time to construct graph: 243.615694459s
Analyzing centrality...
Time to analyze centrality: 665.379783375s
Top 5 Central Nodes: [(10, 4.839701930671995e-5), (31, 4.839701930671995e-5), (35, 3.071658602441506e-6), (29, 2.780508769310654e-6), (30, 2.780508769310654e-6)]
Building predictive model...
Time to build predictive model: 900µs
Prediction completed. Example prediction for the first property: 76.41
Generating visualizations...
Time to generate visualizations: 29.500709ms

Summary of Analysis:
---------------------
Total Properties Analyzed: 73742
Total Nodes in Graph: 73023
Total Edges in Graph: 350666985
Top 5 Most Connected Locations (Centrality): [(10, 4.839701930671995e-5), (31, 4.839701930671995e-5), (35, 3.071658602441506e-6), (29, 2.780508769310654e-6), (30, 2.78
0508769310654e-6)]
```

The output from the project provides a clear and concise breakdown of the analysis process and results. The dataset, which contains 73,742 rental properties, was first cleaned and processed to extract relevant fields such as latitude, longitude, rent, and area. The preprocessing step ensured that only complete and valid data entered the pipeline, which allowed the program to efficiently handle the large volume of records. This step was completed in just 170 milliseconds on my laptop.

Next, the spatial graph was constructed using latitude and longitude coordinates. Each property became a node, and edges were created between nodes if the properties were within 10 kilometers of each other calculated using the haversine formula. This step resulted in a graph with 73,023 nodes and an impressive 350,666,985 edges representing a dense network of geographically connected properties.

The closeness centrality analysis identified the most connected properties in the graph, even those that are close to all other properties in terms of graph distance. Centrality was calculated using breadth-first search to optimize traversal across such a large graph. The top 5 nodes with the highest centrality scores are reported as:

[(10, 4.8397e-5), (31, 4.8397e-5), (35, 3.0716e-6), (29, 2.7805e-6), (30, 2.7805e-6)]
Nodes 10 and 31 stood out as the most connected properties, indicating their potential locations in urban hubs with high accessibility, such as The Marina or Al Reem Island. These areas likely have strong infrastructure and a concentration of amenities, making them desirable in the rental market.

Finally, the predictive model provided quick insights into relative demand or market saturation based on rent per square foot and property size. The model completed its predictions in just 900 microseconds in my laptop, with an example prediction of 76.41 for the first property. The visualizations including centrality maps and regression plots were generated efficiently in under 30 milliseconds. Overall, the output delivers a comprehensive analysis of Dubai’s rental market, uncovering well-connected neighborhoods and highlighting trends that align with the project’s purpose.