Final Project - 125 Points

The required submission artifacts for your final project are three things:

1) Final Project Document. (65 Points)

Your document is a write-up similar to past homework assignments where you should

clearly explain what you did and how you did it. Code snippets should be shared,

commands should be listed, screen shots should be shown. Write your document as if

you were writing an instruction manual for someone trying to replicate your work. Be very

detailed. Your document should also include:

**● overview of what it is your project does (your proposal is probably a great**

**reference for this).**

My project involves developing a data pipeline using MongoDB to store and analyze real estate data. The pipeline will collect data from various sources, clean and transform it, and store it in a MongoDB database. The data will then be analyzed to provide insights into trends and patterns in the real estate market. This will involve performing queries and aggregations on the data using MongoDB's aggregation framework.

These queries/pipeline can extract meaningful insights, such as the average price of active listings and the number of sold listings. This data can then be used by real estate professionals to make informed decisions about pricing, marketing, and buying/selling properties.

● Answers to the following questions:

**○ How will this database scale as it grows?**

I am using MongoDB for my project, which is a highly scalable NoSQL database. MongoDB is designed to be horizontally scalable, meaning that it can handle increasing amounts of data by adding more nodes to a cluster. This allows for easy scaling as the amount of data grows.

In addition, MongoDB uses sharding to distribute data across multiple servers in a cluster, which further enhances scalability.

**○ As more data is inserted over time, how will your system handle the**

**growth?**

We can add more nodes to the cluster, which will allow us to scale the database. Hence, MongoDB is the highly scalable solution for handling the growth as the data amount will increase overtime.

● For a NoSQL database it can be a bit trickier since we do not have a standard

ER model to follow, and the schema is flexible. However, diagram key entities,

attributes and relationships to the best of your knowledge.

1. Key Entities and its attributes –
   1. **Listings** – id, property type, property address, property description, days on market and list price etc.
   2. **Users** – User id, name, email, ph no etc.
   3. **Agent** – id, name, email, ph no, agency name etc.
   4. **National Monthly Inventory** – id, month, country, zip, listing price, active listing, days on market, new listing count, pending listing count, price, price per sq ft, total listing count, pending ratio, etc.
   5. **Market Hotness by County** – id, month, hotness rank, country, county, supply score, hotness score, days on market, unique views per property, demand score, listing price etc.
2. Relationship between entities-
   1. User-Agent Relationship: Many-to-Many Relation. A user can be associated with multiple agents, and an agent can work with multiple users.
   2. Agent-Listing Relationship: One-to-Many Relation. An agent can have multiple listings.
   3. Listing-Property Relationship: One-to-One Relation. A listing has details of a property, such as address, number of bedrooms and bathrooms, square footage, and price.
   4. National Monthly Inventory and Market Hotness County Relationship: one-to-many relationship, as one specific month and location can have multiple market statistics data.
   5. Listing-Showing Relationship: One-to-many, as a single listing can have multiple showings.

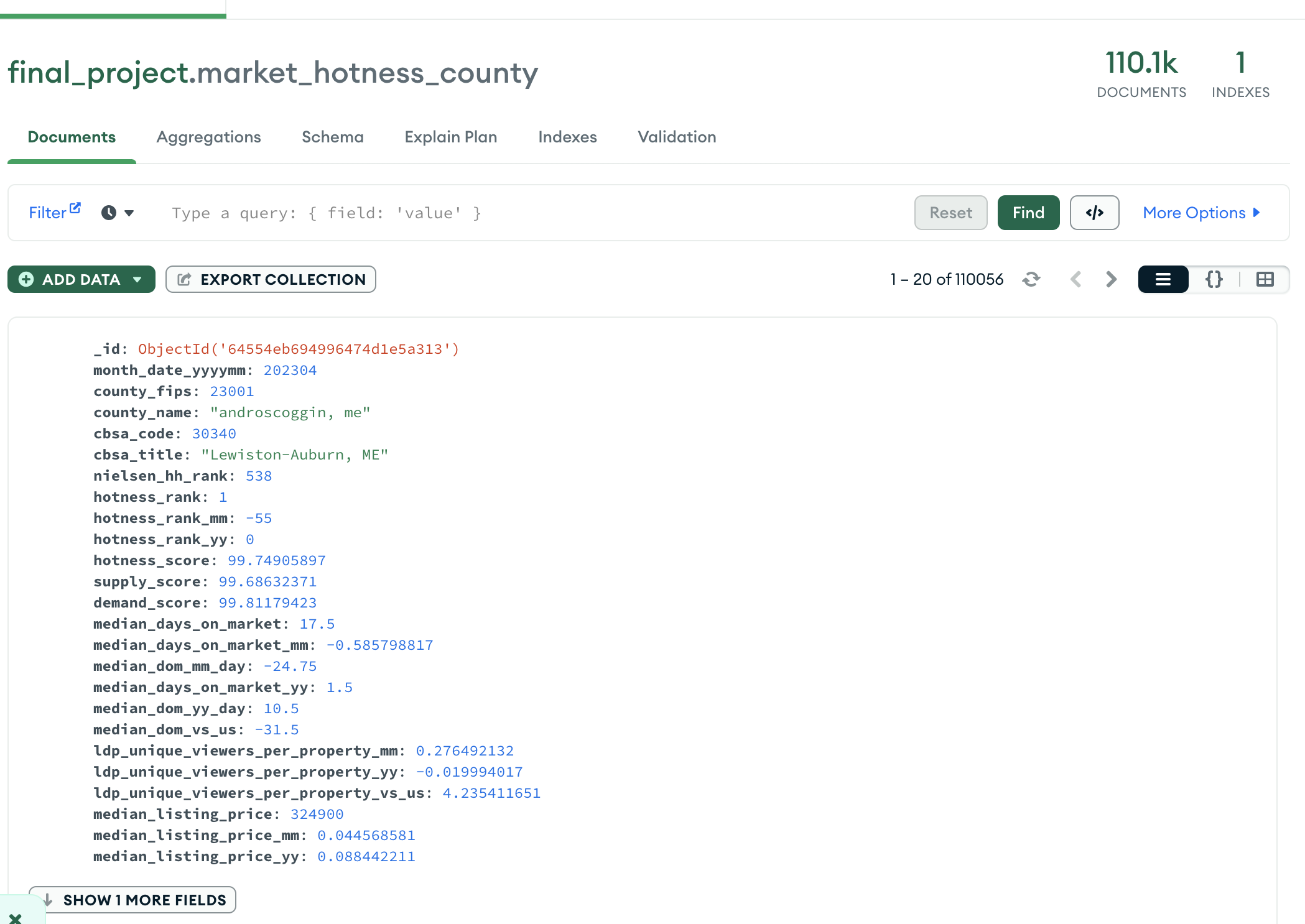
● At least five queries against your data. We are looking for meaningful queries that

demonstrate the purpose of your database. Think in terms of what kinds of

questions you would like to ask against the information you have stored.

Steps to build the database-

1. Download the csv file from the <https://www.realtor.com/research/data/> and imported that using compass.
2. Below is the screenshot of the available tables-

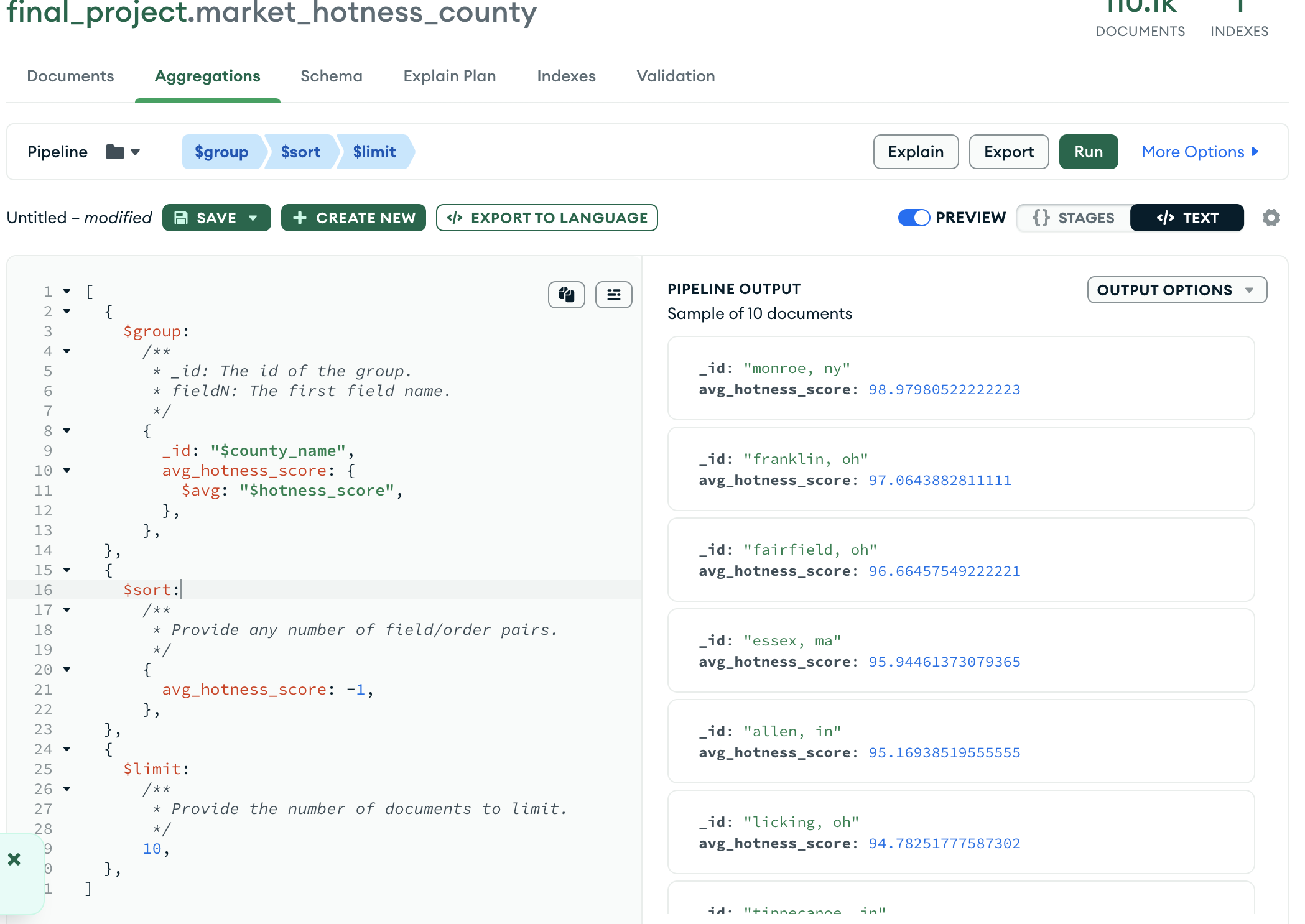


A close-up of a green rectangle

Description automatically generated with low confidence

## Meaningful queries-

1. Which top 10 cities or zip codes have the highest Market Hotness Index scores use this data to identify areas that are currently in high demand or where it may be easier to sell a property.



1. A any seasonal trends in the Market Hotness Index scores? compare the scores across different months or years to identify any trends.

Pipeline code-   
*[*

*{*

*'$group': {*

*'\_id': {*

*'month\_date\_yyyymm': '$month\_date\_yyyymm'*

*},*

*'avg\_hotness\_score': {*

*'$avg': '$hotness\_score'*

*}*

*}*

*}, {*

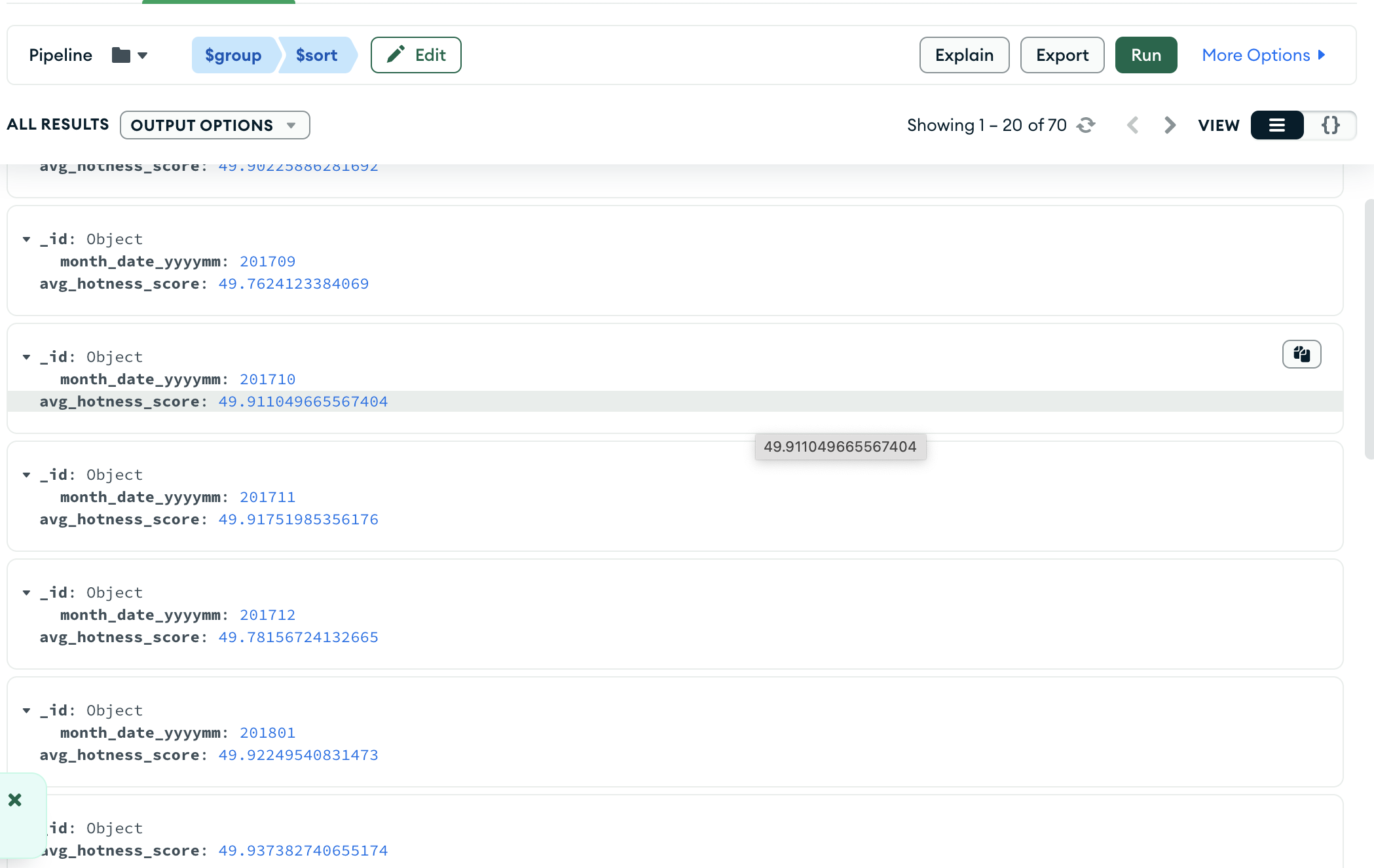
*'$sort': {*

*'\_id.month\_date\_yyyymm': 1*

*}*

*}*

*]*



1. average number of days on market and active listing count for each month in the dataset. (dataset used - National\_historic\_data)

code - *[*

*{*

*$addFields: {*

*days\_on\_market\_bucket: {*

*$switch: {*

*branches: [*

*{ case: { $gte: ["$median\_days\_on\_market", 80] }, then: "more\_than\_80\_days" },*

*{ case: { $gte: ["$median\_days\_on\_market", 40] }, then: "40\_to\_80\_days" },*

*{ case: { $gte: ["$median\_days\_on\_market", 20] }, then: "20\_to\_40\_days" },*

*{ case: { $gte: ["$median\_days\_on\_market", 10] }, then: "10\_to\_20\_days" },*

*{ case: { $lt: ["$median\_days\_on\_market", 10] }, then: "less\_than\_10\_days" }*

*],*

*default: "unknown"*

*}}}*

*},*

*{*

*$group: {*

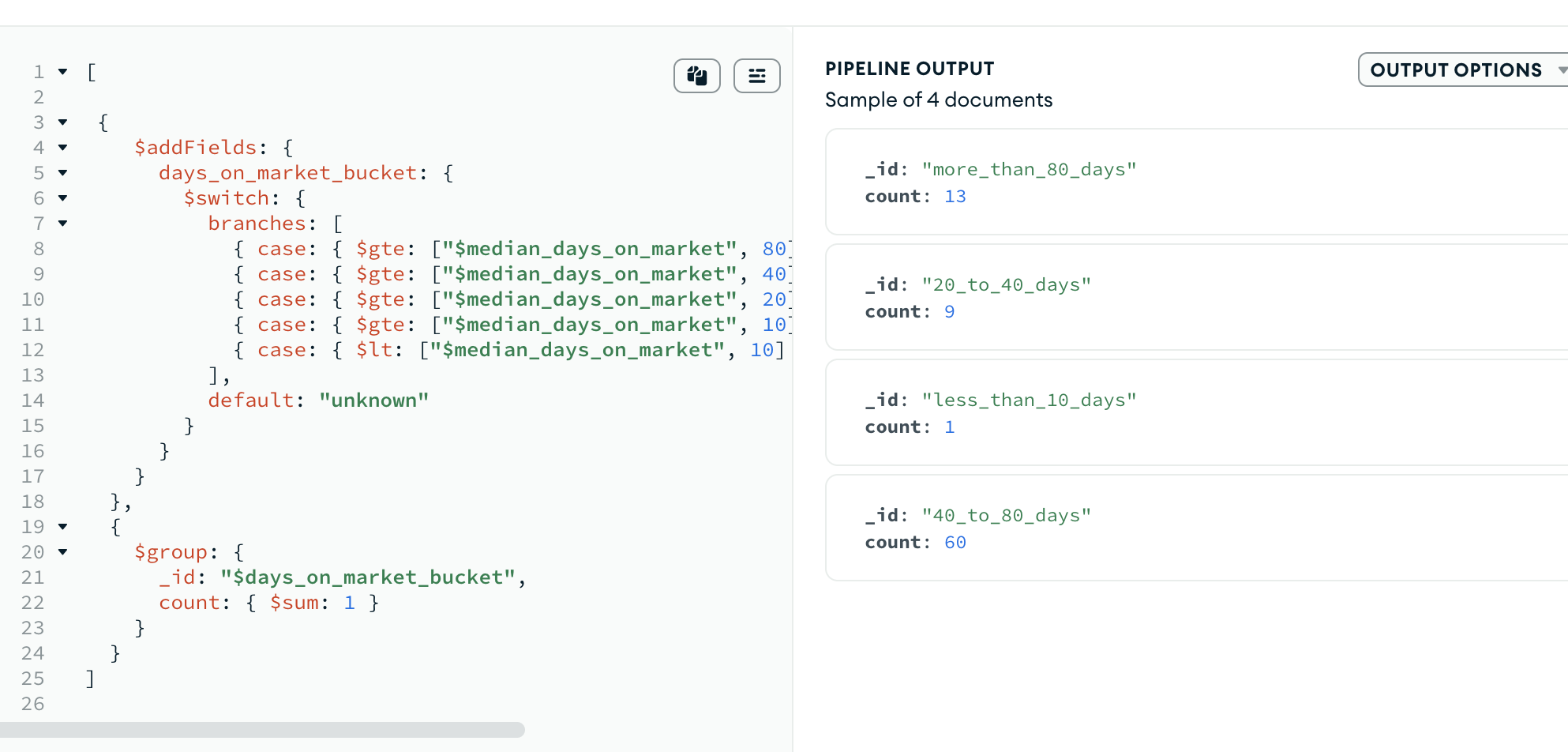
*\_id: "$days\_on\_market\_bucket",*

*count: { $sum: 1 }*

*}*

*}*

*]*



1. What is the total number of listings per month in 2023 with a median square footage for properties in the United States?

code –

*[*

*{*

*$match: {*

*"country": "United States",*

*}*

*},*

*{*

*$addFields:*

*{*

*'year': {*

*'$toInt':*

*{*

*$substr: ['$month\_date\_yyyymm', 0, 4]*

*}*

*},*

*'month': {*

*$substr: ["$month\_date\_yyyymm", 4, 2]*

*},*

*'medium\_sq\_foot': '$median\_square\_feet'*

*}*

*},*

*{*

*$match:*

*{*

*'year': {'$eq': 2023}*

*}*

*},*

*{*

*$group:*

*{*

*\_id:*

*{*

*year: '$year',*

*medium\_sq\_foot: '$medium\_sq\_foot'*

*},*

*no\_of\_listings: {$sum: '$total\_listing\_count'}*

*}*

*},*

*{*

*'$sort': {'\_id.year': 1}*

*}*

*]*

A screenshot of a computer

Description automatically generated with medium confidence

1. calculate the no of active listings and their avg price. Also calculate the total sol listings. Use monthly national data.

Code –

*[*

*{*

*$match: {*

*country: "United States",*

*month\_date\_yyyymm: 202304*

*}*

*},*

*{*

*$group: {*

*\_id: "$country",*

*active\_listing\_count: { $sum: "$active\_listing\_count" },*

*avg\_listing\_price: { $avg: "$median\_listing\_price" },*

*pending\_listing\_count: { $sum: "$pending\_listing\_count" },*

*total\_listing\_count: { $sum: "$total\_listing\_count" }*

*}*

*},*

*{*

*$project: {*

*\_id: 0,*

*active\_listing\_count: 1,*

*avg\_listing\_price: 1,*

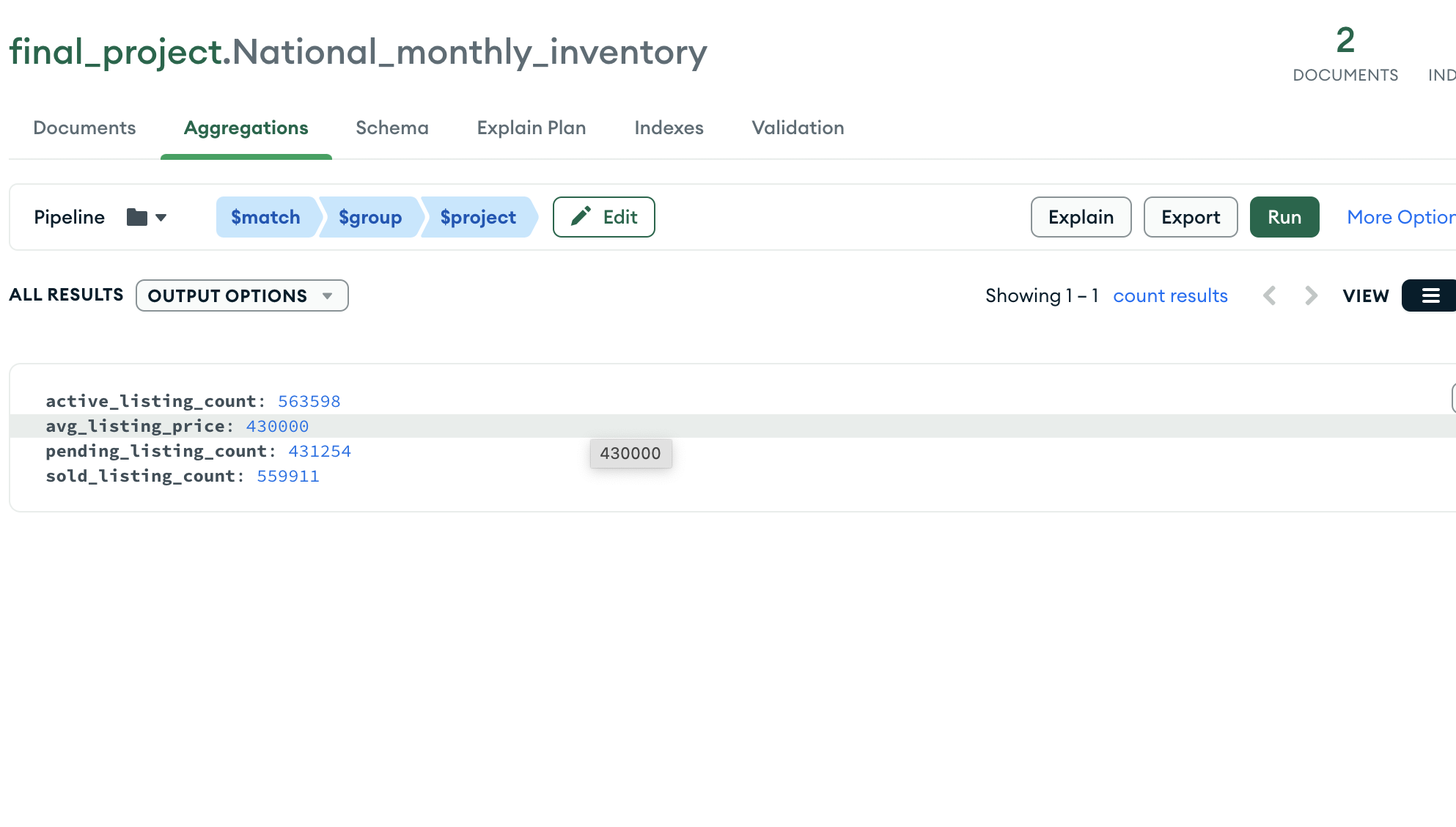
*pending\_listing\_count: 1,*

*sold\_listing\_count: { $subtract: [ "$total\_listing\_count", "$pending\_listing\_count" ] }*

*}*

*}*

*]*

**

Your document must also contain links to the following video:

2) 3-5 minute YouTube Presentation. (60 Points)

A short summary of your project, and results. Summary videos will be shared with the

class on the final day of class, and students should make an effort to be present for the

live lecture that day as there will be Q&A after each video. Important: You will lose

points if you are under three minutes, or over five minutes.

Youtube link - <https://youtu.be/dcmKUewSAu8>

Summary-

This project is required to create a centralized source of information for the real estate market. I am using Mongo DB to store the data as it is highly scalable database which can handle the increasing amount of the data overtime. I will be creating the pipelines to clean and transform the data which can be used to provide the insights of trends and pattern in the real estate market. I have calculated some key metrics in this project which are - identifying top 10 cities or zip codes with the highest Market Hotness Index scores, identifying seasonal trends in the Market Hotness Index scores, and calculating the average number of days on the market and active listing count for each month in the dataset. This data can also be used by real estate professionals to make informed decisions about pricing, marketing, and buying/selling properties.

I have used Realor.com dataset in this project, which are available at this website by downloading the csv files. To import this data into MongoDB, I have used the MongoDB Compass tool. You can click on this + sign here on left side to create the database and then create the collection and import the csv file by clicking on import data.