CIS3200 Term Project Tutorial

Authors: NATHAN KWOK,AN NGUYEN,TYLER TRINH,HANS ERON VALDERAMA

Instructor: Jongwook Woo

Date:

Lab Tutorial

COVID-19 Data Analysis using ElasticSearch

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Get data manually using REST API
- Upload a data set to Elasticsearch
- Create a Data View in Kibana
- Create a Visualization
- Create a ML regression model using ES

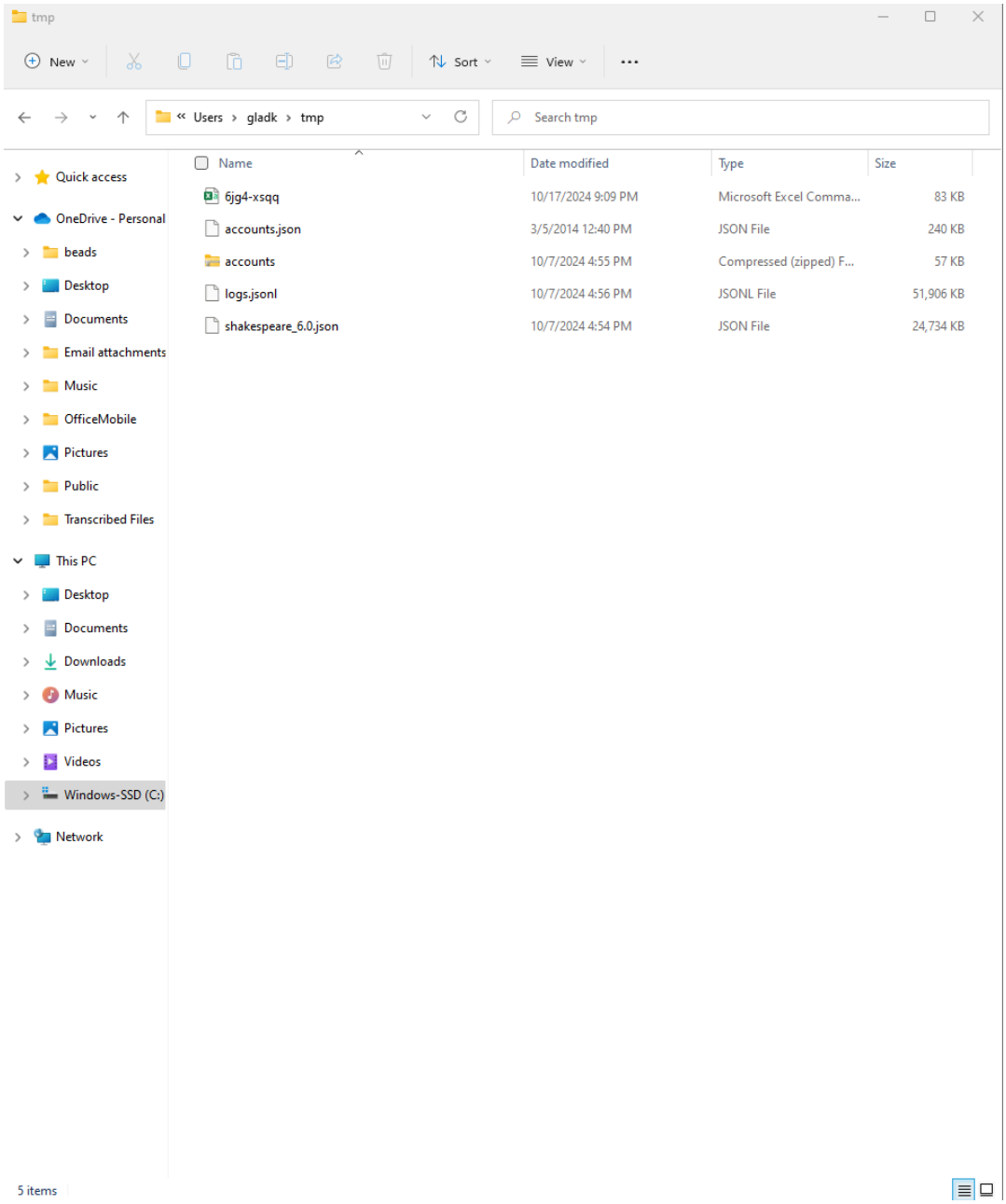
Step 1: Get data using Git Bash

This step serves to get the data needed and upload the resulting CSV file to Elasticsearch. This step is to get data manually….

1. Open Gitbash.
2. Use cd to go to the desired folder to download the file.
3. Download the file using the curl - command as follows:

```
-  curl -O https://data.cdc.gov/resource/6jg4-xsqq.csv
```

4. After this, you should see the file successfully downloaded in your desired folder.

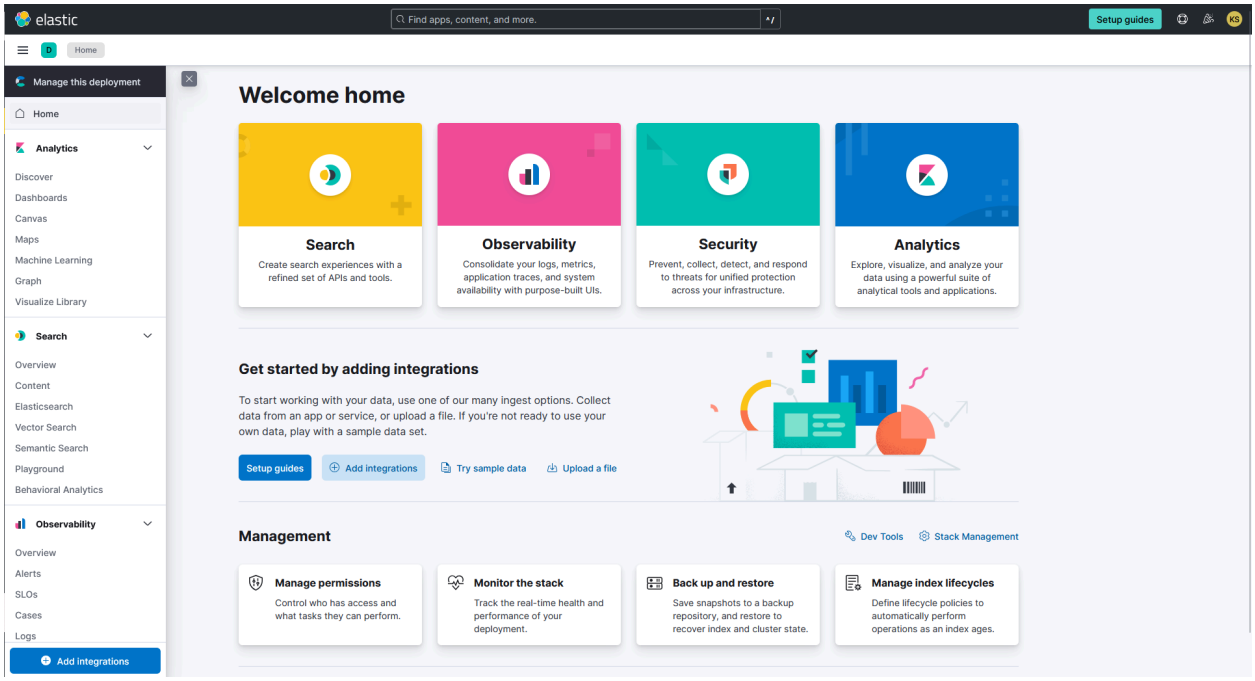


Step 2: Upload the CSV file to ES and Create a Data View

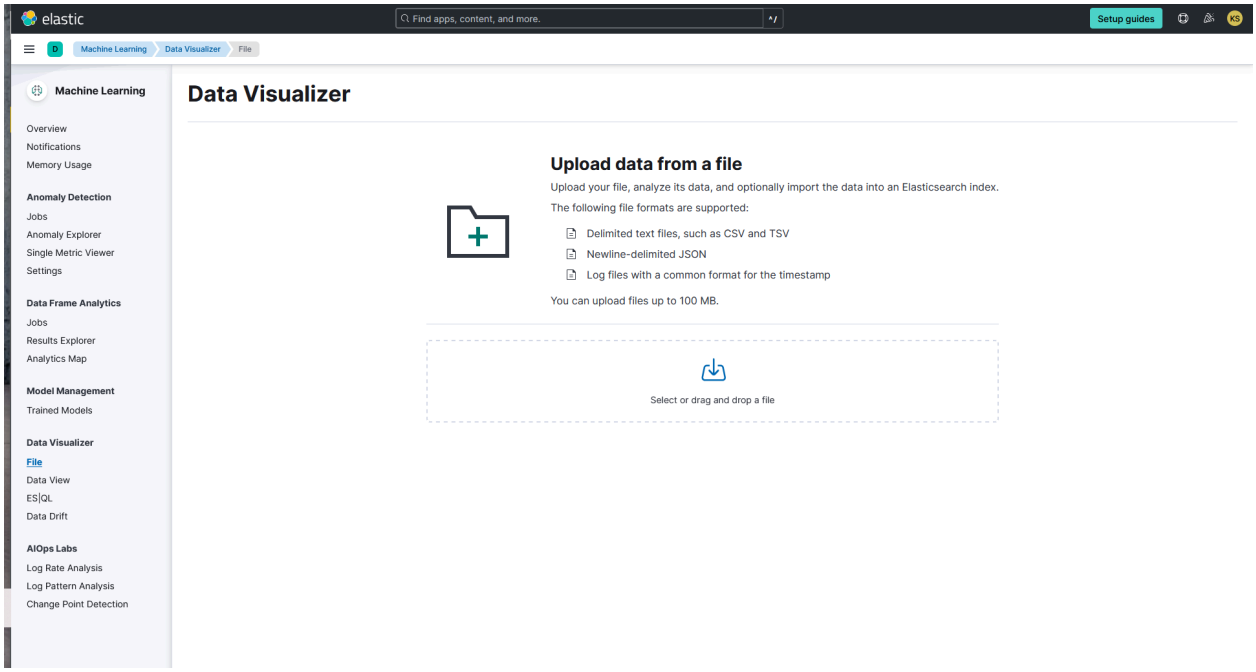
1. Go to https://cloud.elastic.co/login → Choose login with Microsoft → enter your school and email and password and log onto your Elastic Cloud account.

2. Click on your deployment to enter the Kibana homepage.

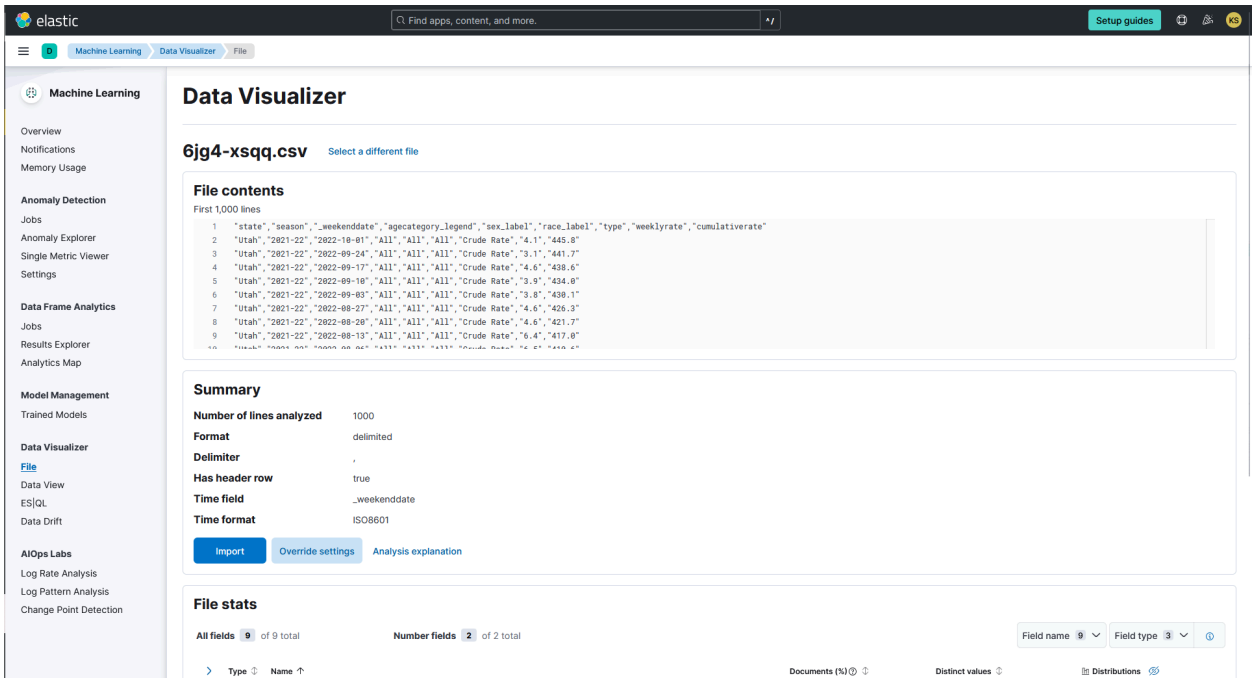
3. Go to the **Analytics** → **Machine Learning** from the Kibana homepage.

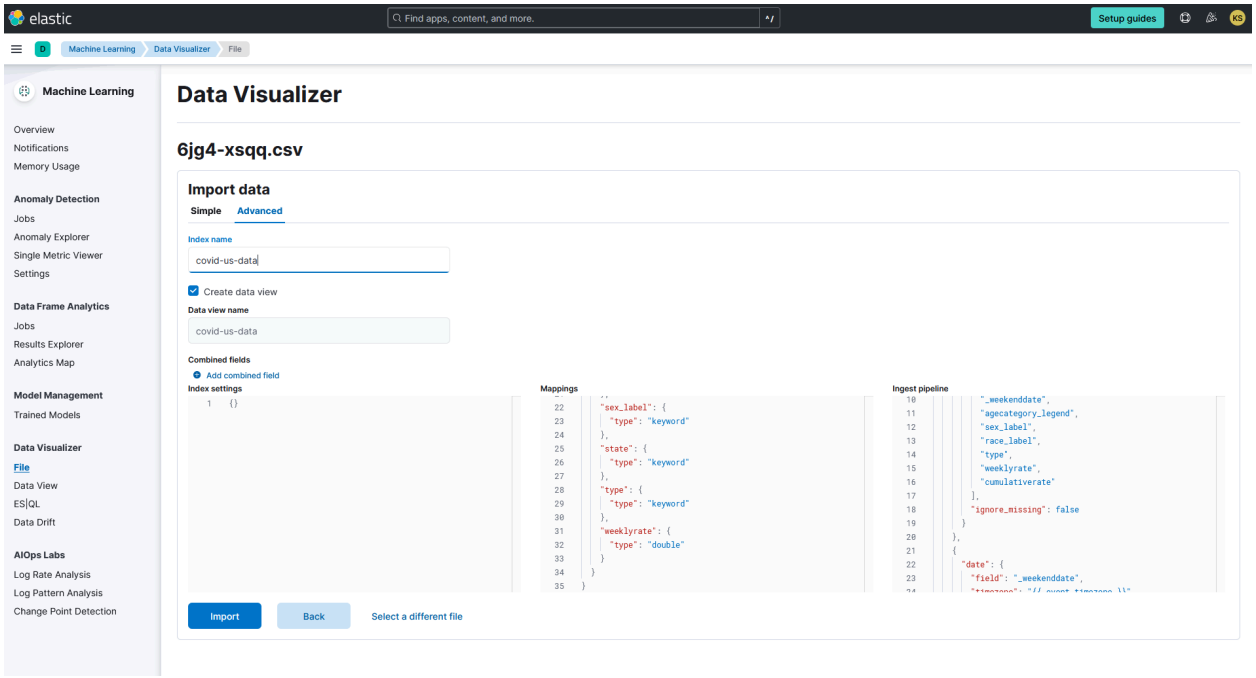
4. 

5. From there, go to **Data Visualizer** → **File**.

6. 

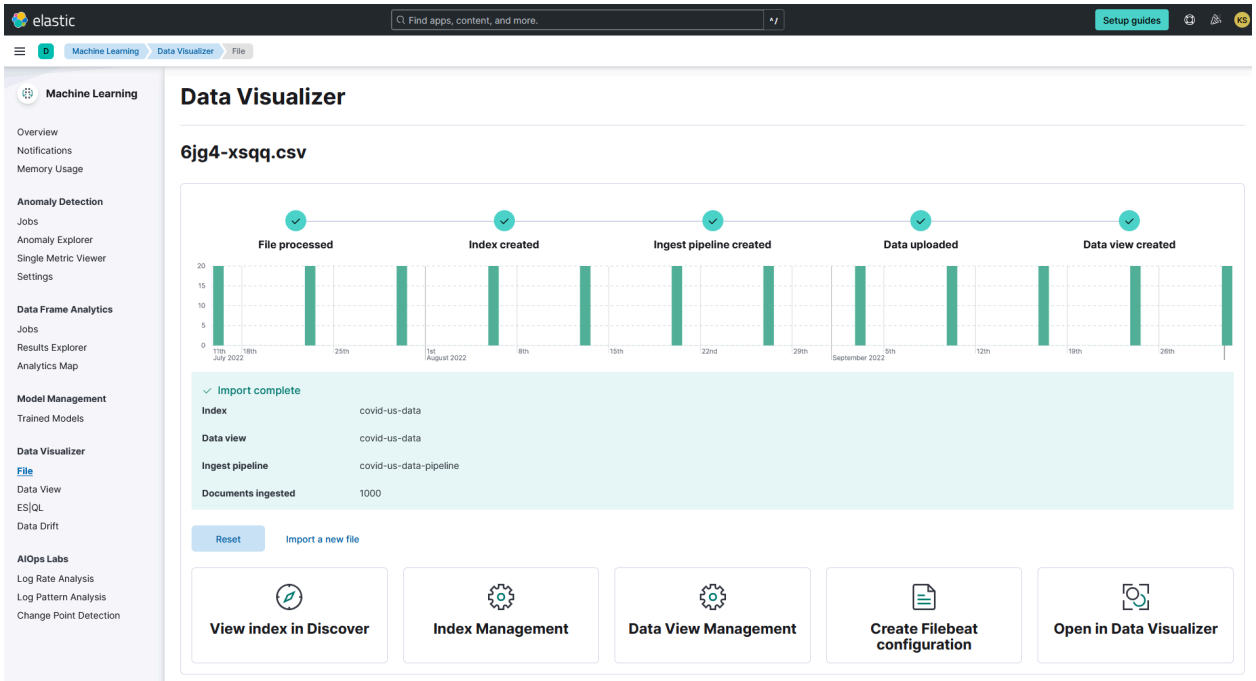
7. Click on the “Select or drag and drop a file option” and select the CSV file downloaded previously. Then click “Import” at the bottom of the screen.

8. 

9. 

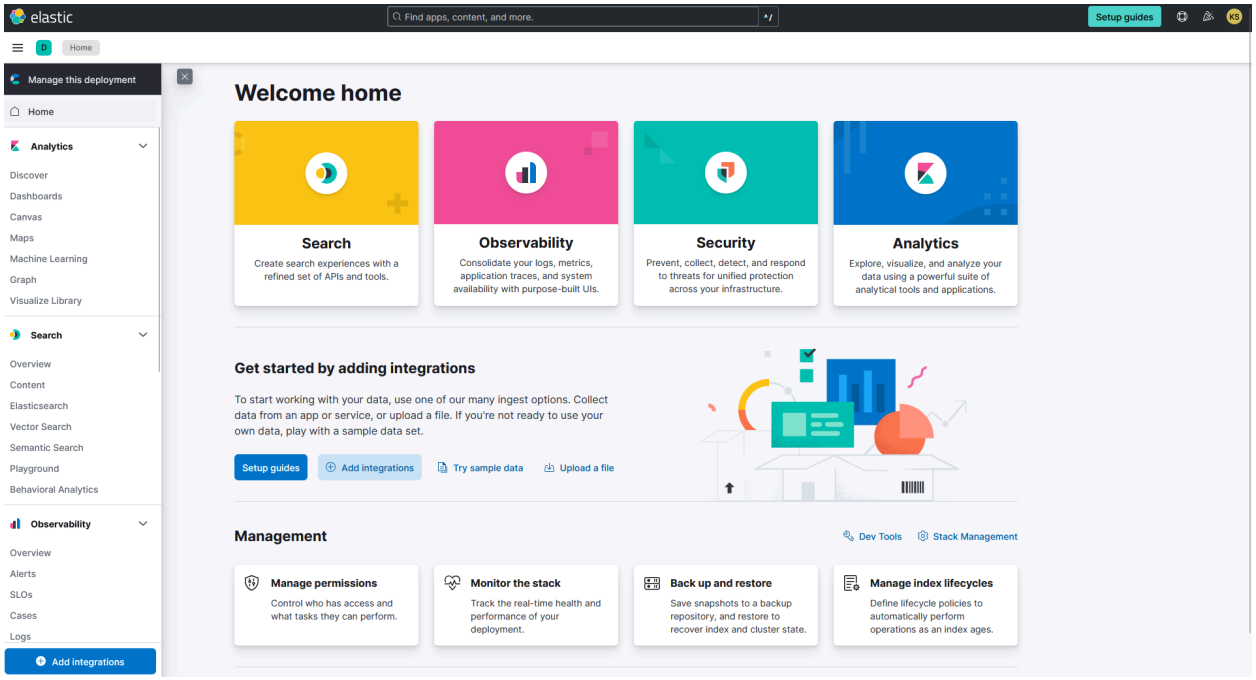
On the next screen, enter “covid-us-data” under “Index name” then click “Import” at the bottom.

Once the data has been successfully imported, you should see the following screen.

10. 

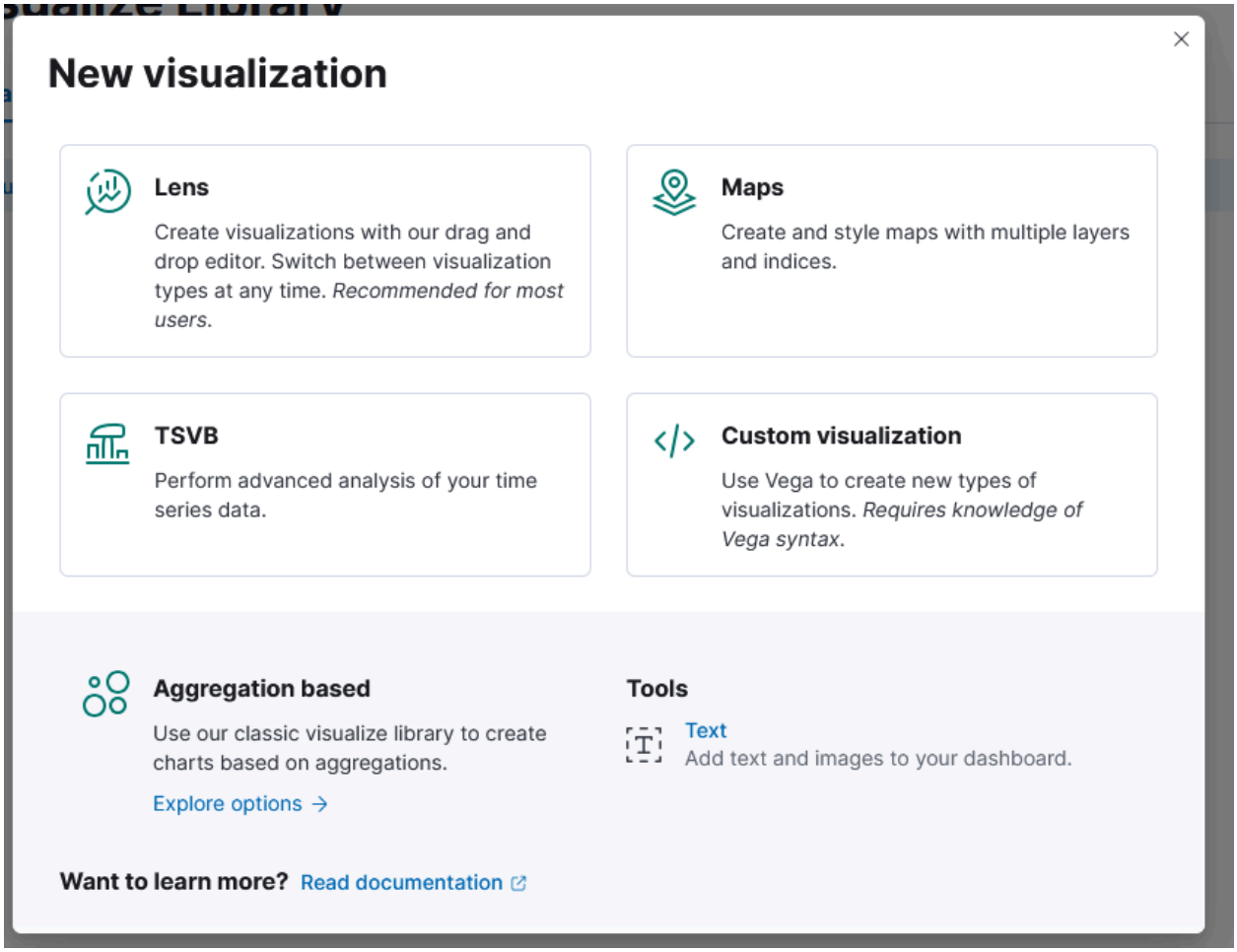
Step 3: Visualization

Explain what this step is for. This step creates a visualization using the data set.



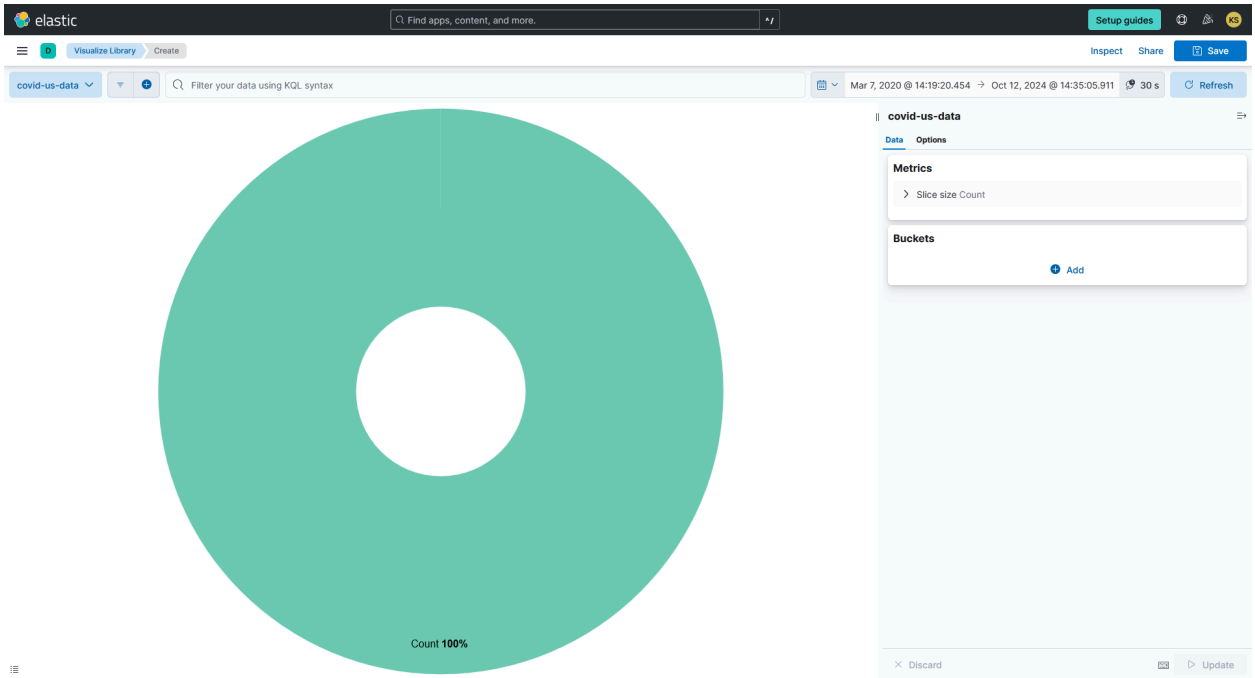
1. On the Kibana homepage, go to Analytics → Visualize Library.

2. Click on “Create Visualization” and select “Aggregation based” then select “Pie” chart and make sure to select “covid-us-data” as the data source.

3. 

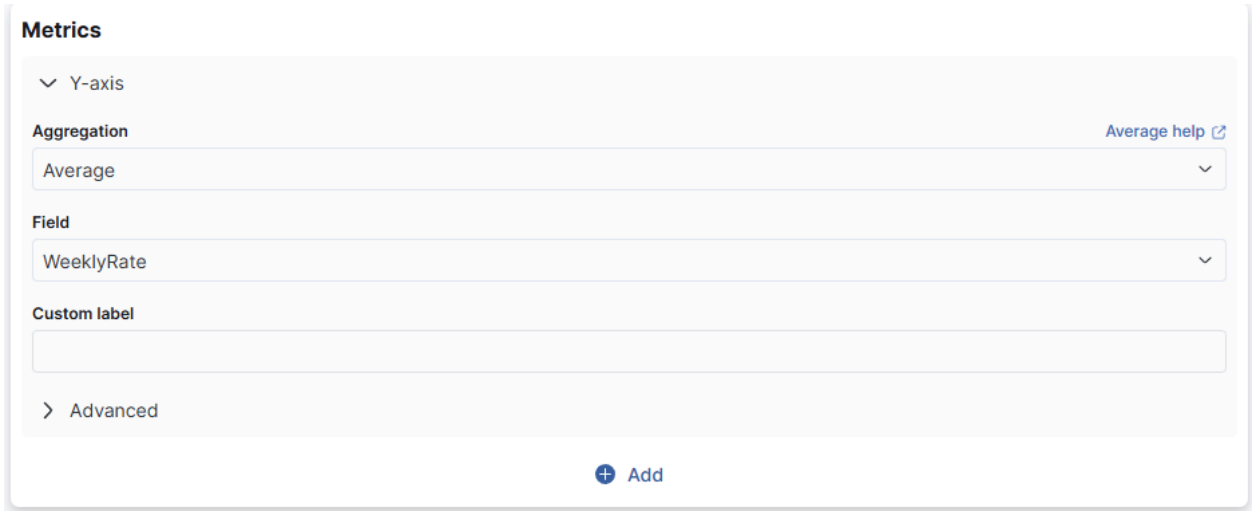
4.At the top, select the date Absolute “3/7/2020” to Absolute “10/12/2024” and then refresh.

Initial screen will be presented:



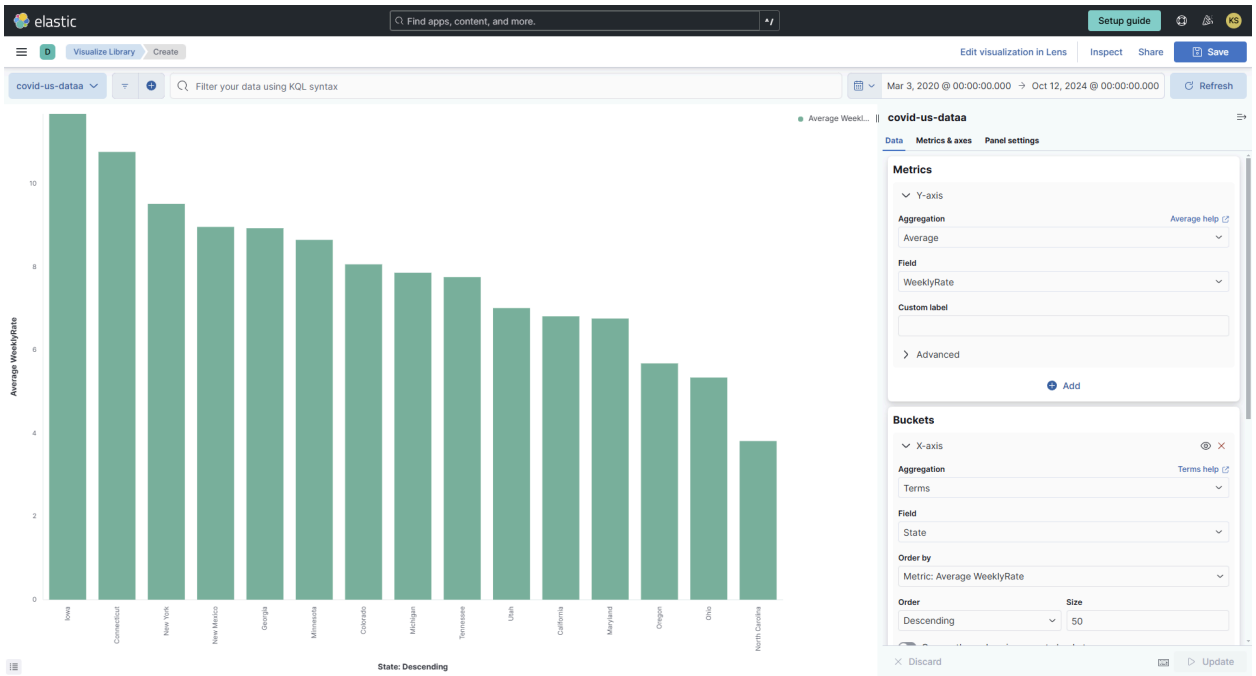
5.Adjust metrics to the states that have the highest rate of weekly cases of covid 19

6.Adjust metrics on Y-Axis to average of weekly rate:



7.Create bucket for the X-Axis that separates into state and aggregate by terms “Average Weekly Rate”

8.After finishing the aggregation you will see the graph that shows the average weekly rate of covid where IOWA is the highest.



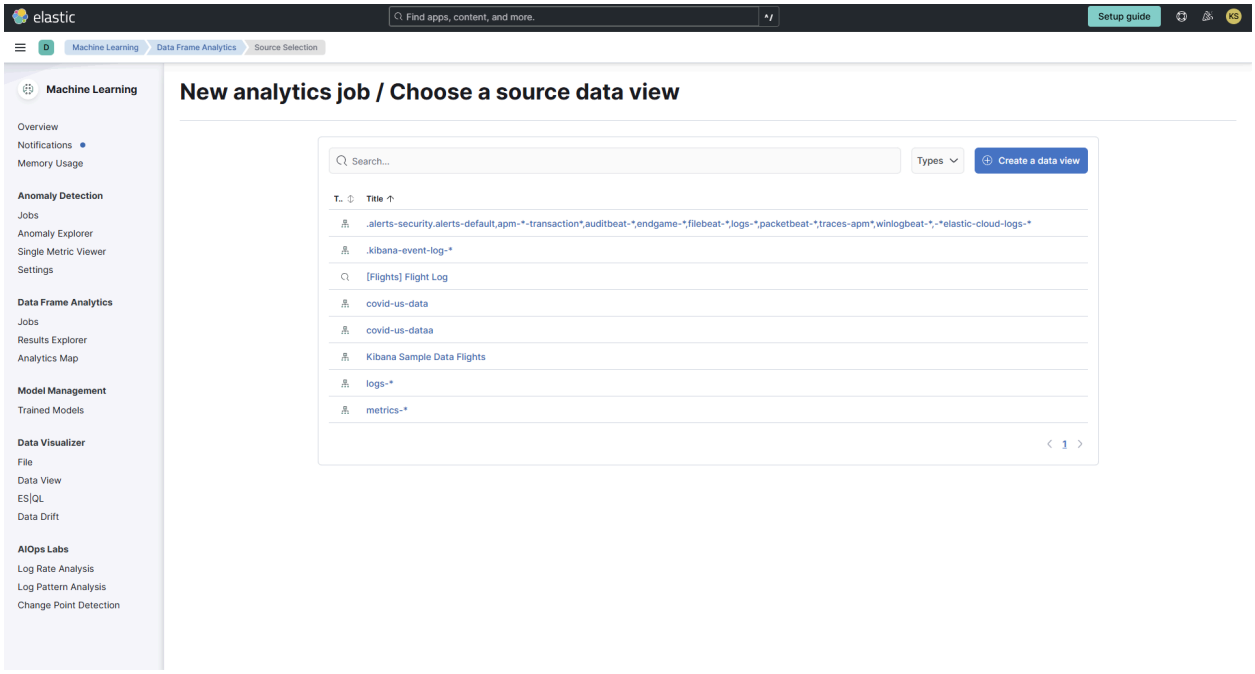
Step 3: Create a Regression Model

This step is to create a regression model within ElasticSearch.

1. To start off, you will need to import the data set into ElasticSearch and create a data view for your data set. You can follow STEPS 1 and 2 to learn how to import data into ElasticSearch and create your own data view.

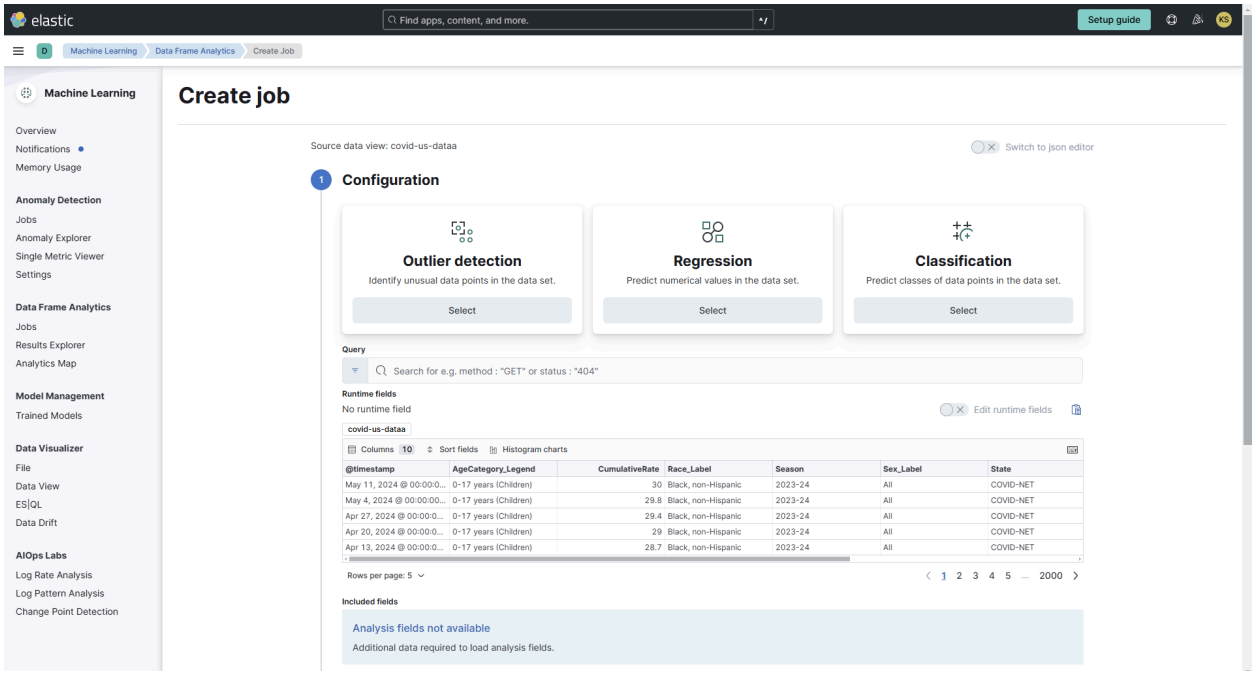
2. Select in Kibana > **Analytics> Machine Learning>Jobs**

3. Click on the button “Create Job”

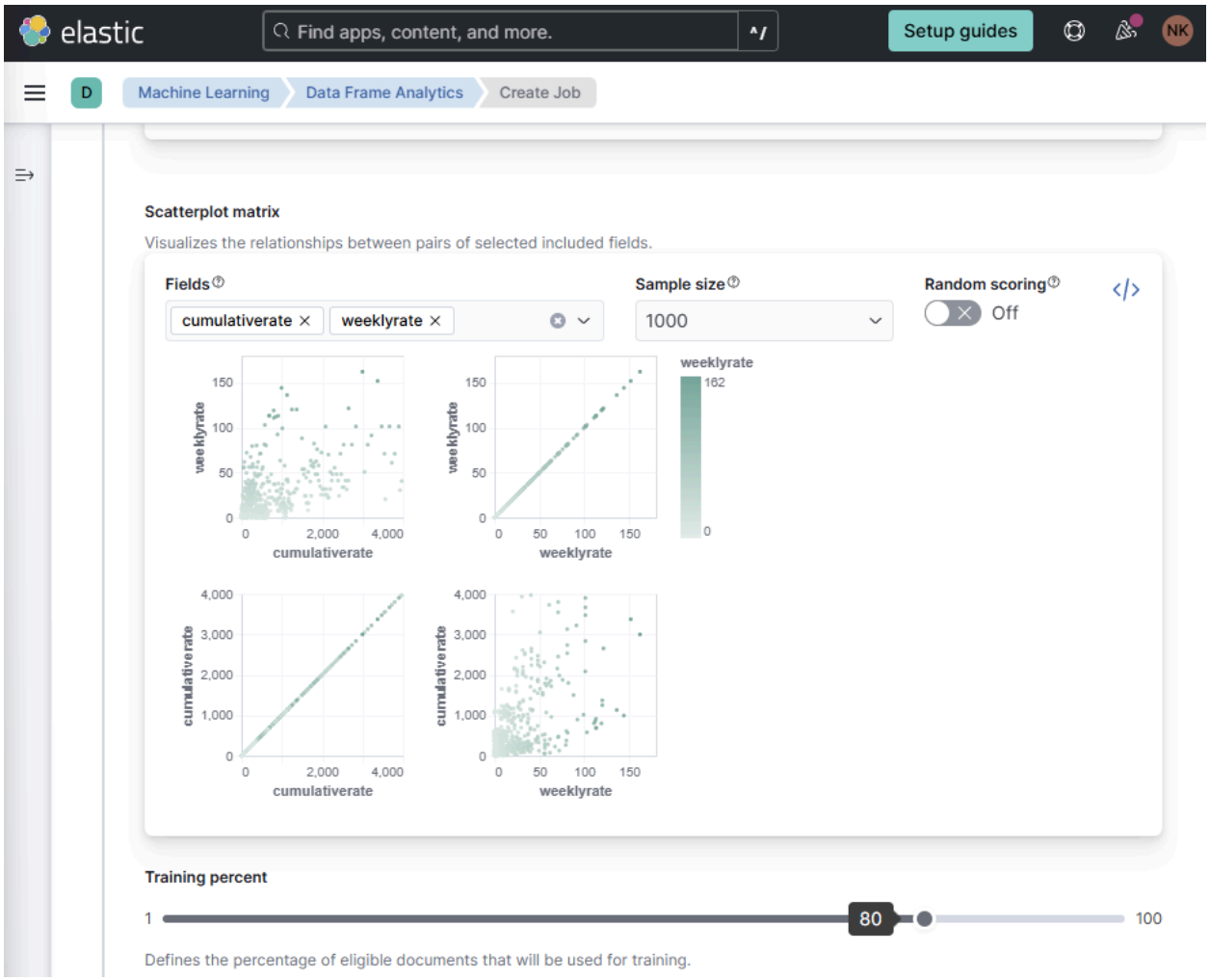
4. 

For selection of data view, click “covid-us-counties” or on the data view name that you’ve created if you decided to name it something else. (In this example we will be covid-us-dataa as our data view name.)

5. Select the regression box in the middle of the 3 choices as that is the model we will be making.

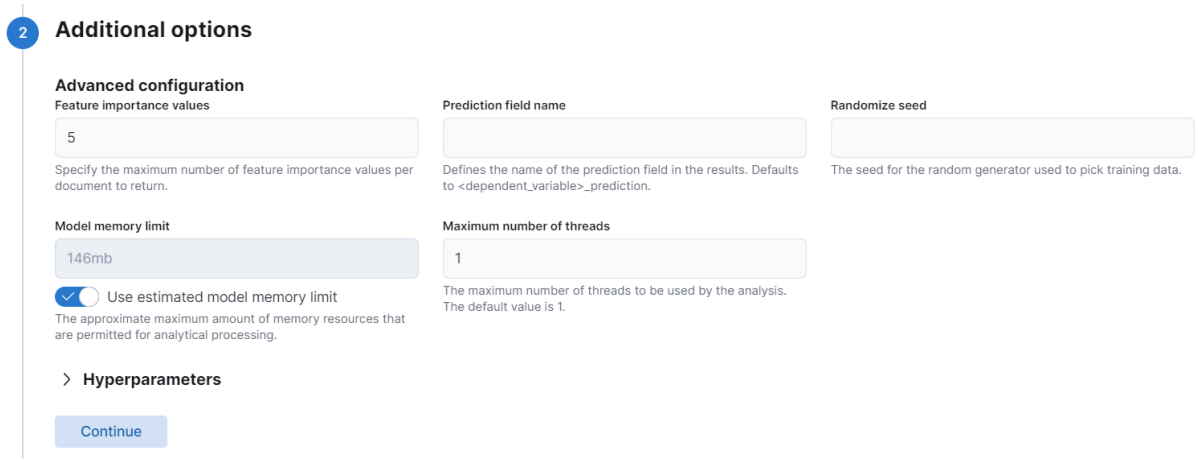
6. 

7. In the query, you may put “weeklyrate> ”. This is because we are doing a regression on weekly and want to filter out weekly rate with an amount of 0.
8. In the dependent variable field, select **weeklyrate** as it is the numeric field we want to predict.

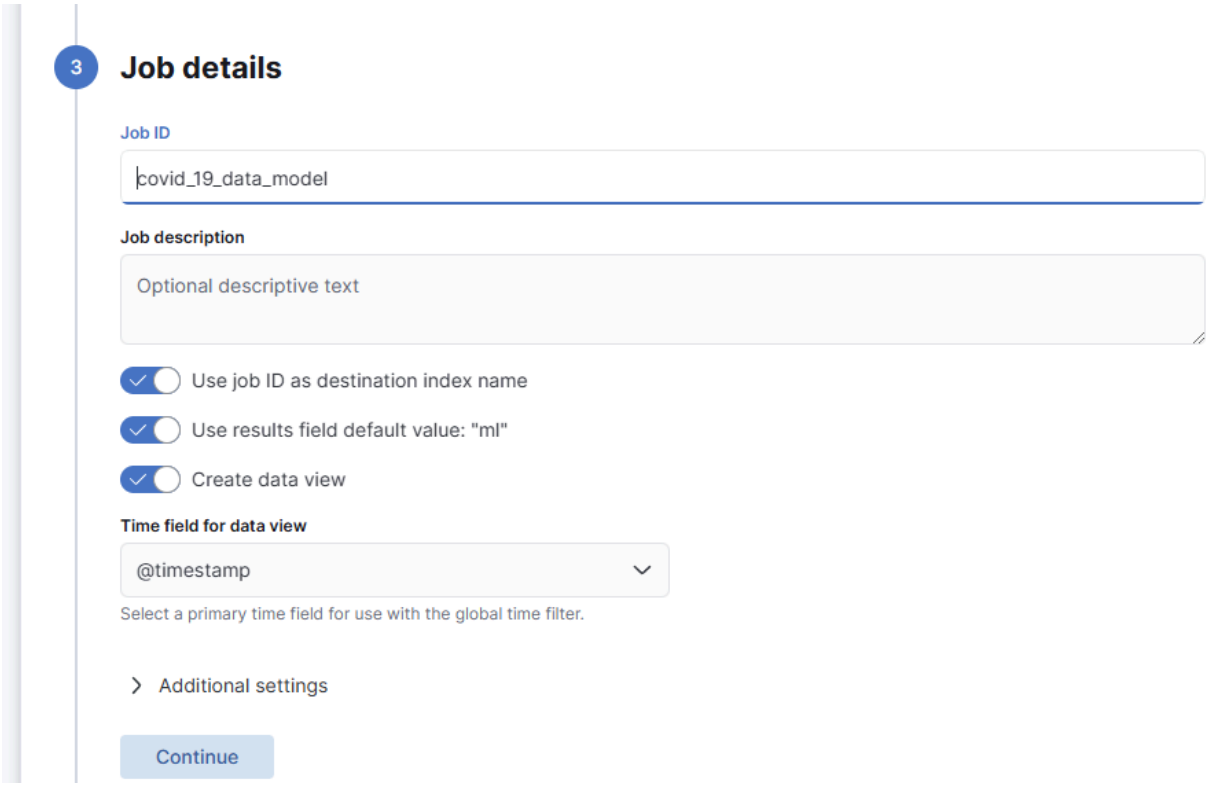
1. For the **training percent** select 80% as it will randomly select 80% of the source data for training. Hit the continue button to move onto the next area.
2. 

3. Set feature importance values to 5.

a. You don't need to change the memory model limit for this section. Hit the continue button to proceed to the next step.

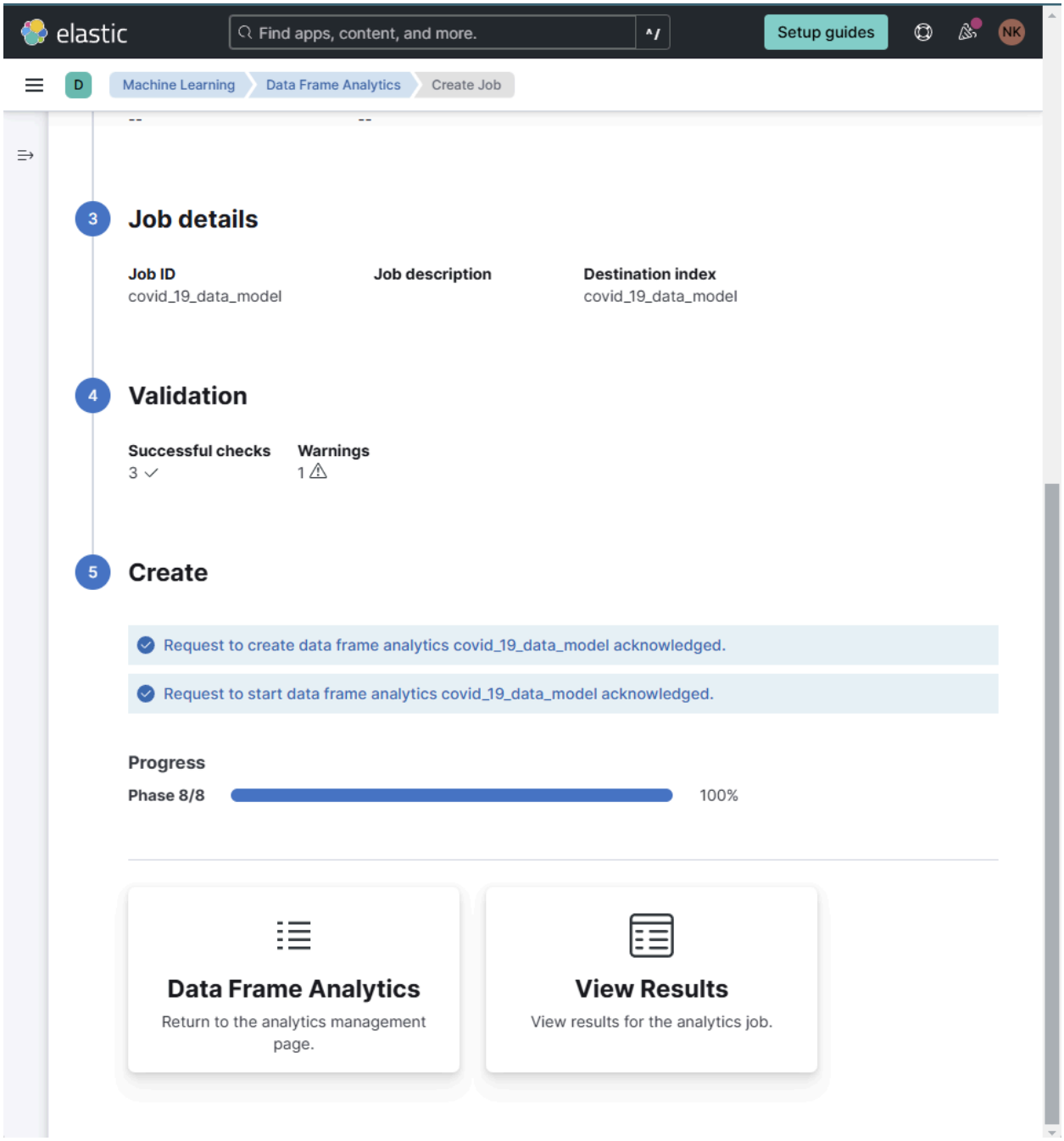


4. For the job ID, you can name it whatever you'd like, but you should make it a name that expresses what the model does like **"covid_19_data_model".**

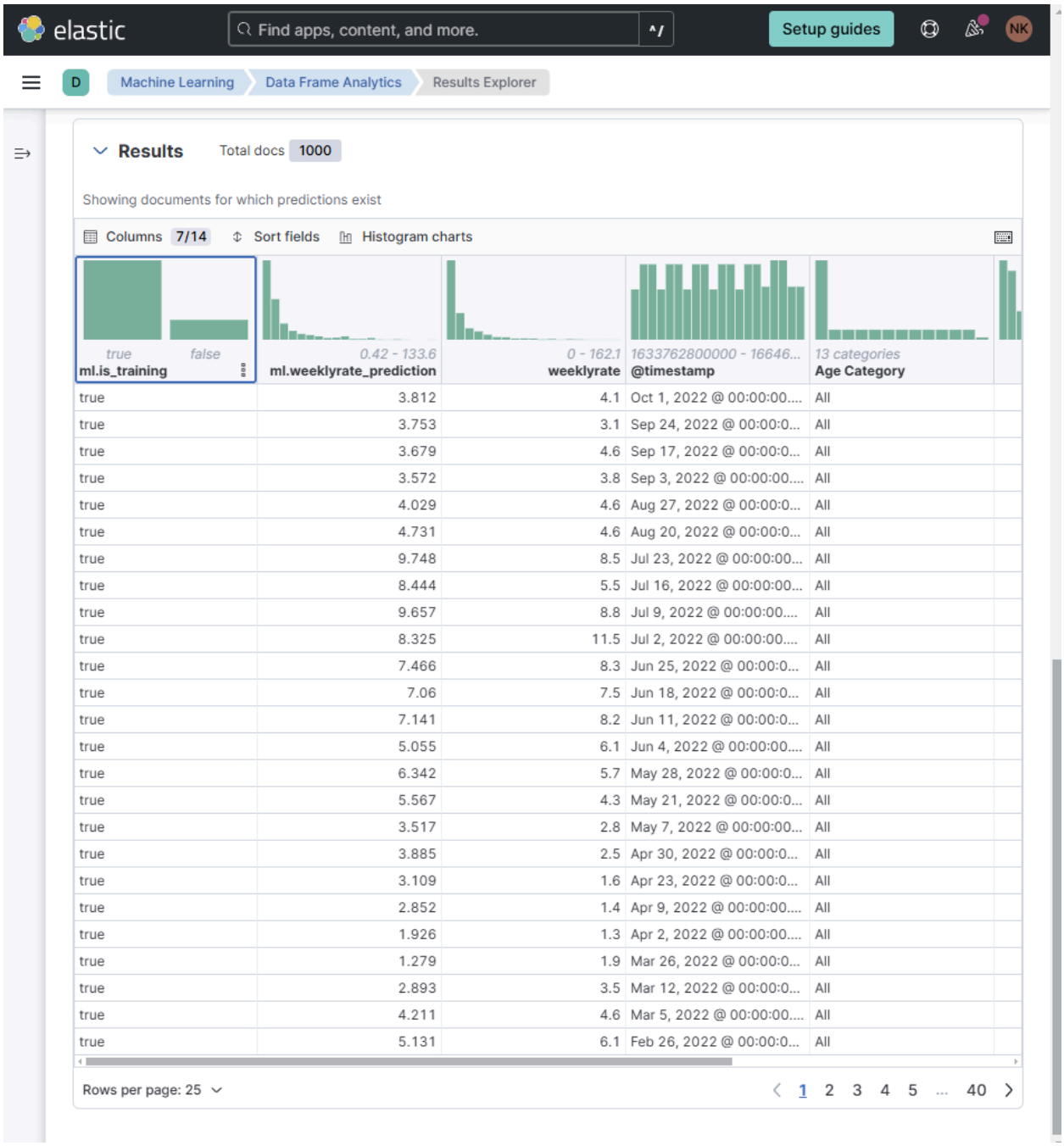
5. 

6. In the **Validation** section, you will find that there are two warnings on training percent and feature importance. This is fine to have and just continue on.
    a. Because you are using such a large amount of data, it takes a while for ElasticSearch to complete training and testing the model.
7. 

8. To see the progress of the regression model, simply click on the **View Results** button to view your newly created model.
9.

10. 

11. The table shows a column for the label, dependent variable(weekly rate),which contains the actual values we are trying to predict. It also shows a column for prediction values(ml.weeklyrate_prediction).
12. The ml.is_training column proves if the document was used in the training or not by stating “true” or “false”

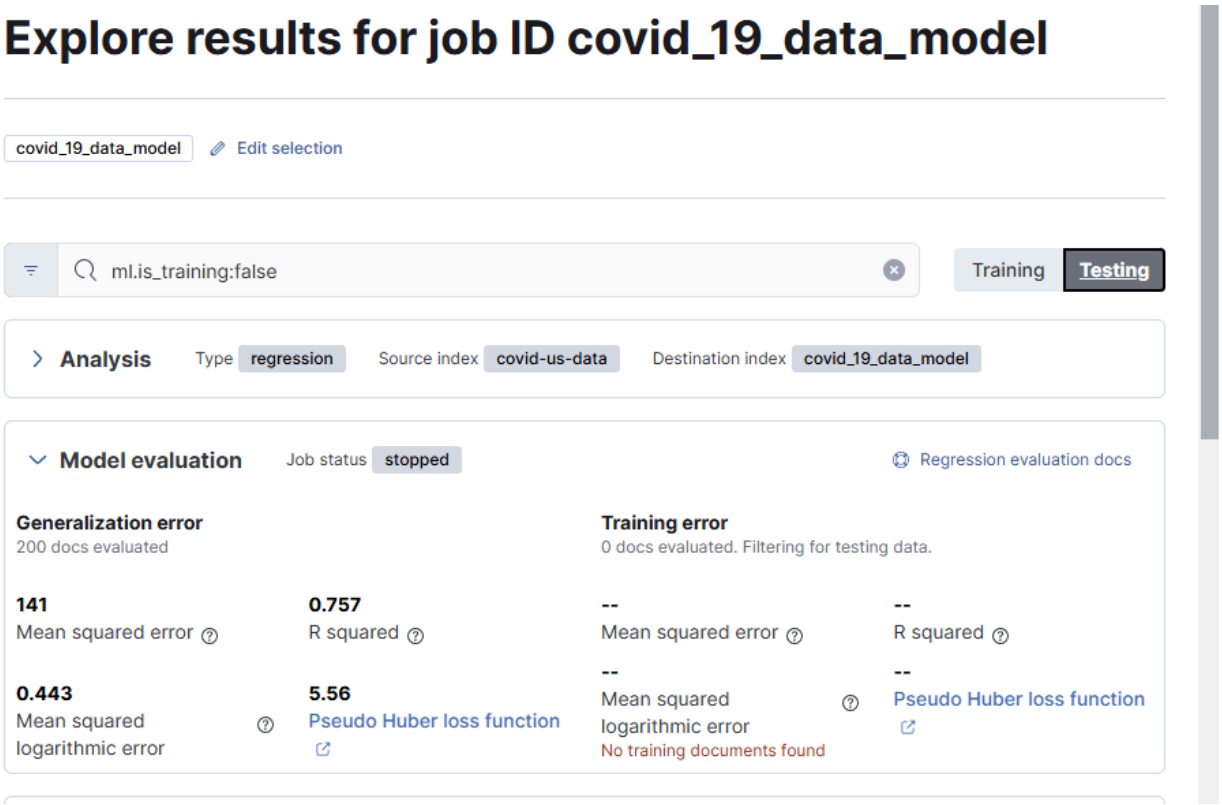
13. 

14. You can see the accuracy of the model with training set data. Select the **Testing** button to view.

    a. Mean Square Error(MSE) is 141. A MSE of zero means that the model predicted the dependent variable (weekly cases) with perfect accuracy. It's highly unlikely a MSE of zero will ever appear.

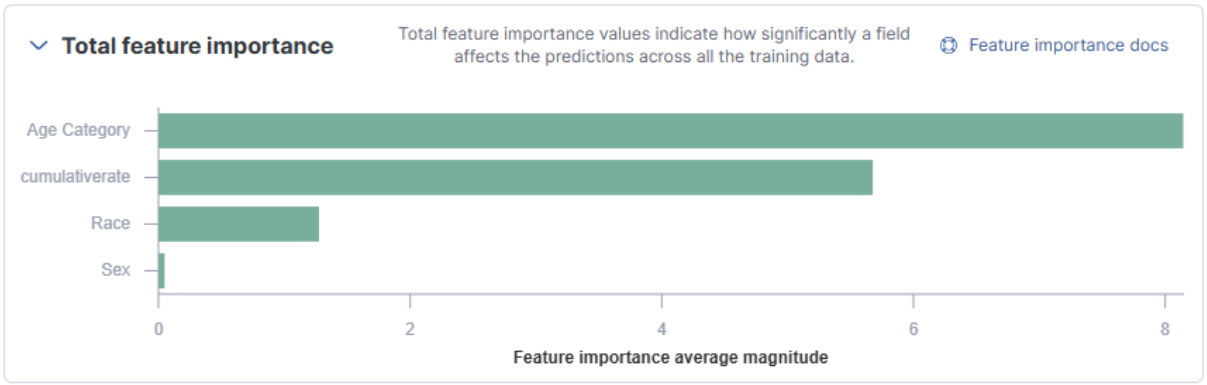
b. R squared is 0.757. The closer to 1 R squared gets, the higher the accuracy. It is also highly unlikely that you will get an R squared of 1. If you do, you may need to rebuild and check your model for any errors.

15.



16. You can see the total feature importance as follows. In this particular model, we see that the field, **Age Category**, had the highest impact on the dependent variable “Weekly Cases”.

17.



THIS IS THE END OF THE TUTORIAL