ITCS 6100 Big Data Analytics for Competitive Advantage

Fall 2022 - Group 14 – TERM PROJECT

Project Deliverable 1: Group Formation and Project Understanding

First Street Foundation (FSF) Flood Risk Summary Statistics

TEAM

Team Members

Harika Moole – 801318334
Noel Vijay – 801274982
Ram Vishal Singh – 801307348
Ritvik Kondabathini – 801275238
Sarath Chillakuru – 801314925

Communication Plan

- **Methods of Communication**: Slack for day-to-day conversation and syncing. Email for a more formal and documented way of communication.
- **Communication response times:** All team members try to communicate at the earliest possible time. Any absence or away-from-work status is intimated in the communication channel.
- **Meeting attendance:** Meeting attendance is not fully mandatory. This implies that it is not necessary for all team members to attend every meeting but at least 3 out of 5 members need to be present. The ones not attending shall inform the rest well in advance. It is appreciated if all can attend anyways.
- **Version control:** All changes by the team members are pushed to the artifact by raising pull requests which are first reviewed by the peers before merging.
- **Division of work:** Work is divided between the team members in the form modules. Any overload can be brought to the notice of the other members so it can be divided or redistributed.

Project artifact repository

All of our work can be found in the public repository that has been created on GitHub.

The link to the repository is - https://github.com/nvijay1/bigdata14

BUSINESS PROBLEM, OPPORTUNITY, DOMAIN KNOWLEDGE

Business Problem

The dataset presents CSV files of flood statistics for the 48 contiguous states at the congressional district, county, and zip code level. The CSV for each of these geographical extents includes statistics on the number of properties at risk according to FEMA, the number of properties at risk according to First Street Foundation.

The flood statistics data can help us understand and gain insights on how well the government or an organization needs to be prepared during floods and helps us to get an estimate of the damages that would incur.

Domain References

https://assets.firststreet.org/uploads/2020/06/first_street_foundation__first_national_flood_risk_assessment.pdf

https://firststreet.org/research-lab/

SELECTION OF DATASET

Dataset

https://registry.opendata.aws/fsf-flood-risk/

RESEARCH OBJECTIVES AND QUESTIONS

Research Objective

Using the historical data from FEMA and First street foundation, we aim to predict the number of properties that would be affected in case of floods with various risk factors. We also intend to study data from the two organizations and provide a comparative analysis of how each organization calculates the number of properties affected with floods of various risk scores.

Example Descriptive Questions

1. Comparing the annual peak flows by regions (for a tenure of 10 years or more)
2. Where are there unexpectedly high building and content loss, given the number of structures?
3. Identifying areas of high potential losses.
4. What are the historic flood outlines for each state?

Example Predictive Questions

1. We will forecast the likelihood of a 100-year flood in each state in any given year?
2. Find the zip codes with the highest percentage of properties at risk of flooding in the future.
3. Find the counties in each state that are most at risk.

DELIVERABLE - 2

The Flood risk summary dataset is divided into four parts providing summary statistics correlating to each state, district, county, and zip code. Four files include the data collected by First street foundation and FEMA.

1. Fsf_flood_zcta_summary.csv
2. Fsf_flood_county_summary.csv
3. Fsf_flood_cd_summary.csv
4. Fsf_flood_state_summary.csv

We have utilized state, county, and district data for data visualization and as a part of data modeling we have utilized the zip code data. The dataset includes the number of properties damaged by floods with various flood risk factors ranging from 1-10. The data also includes the number of First Street properties with flooding in 30 years in the Return Period 500 scenario. A 30-year flood does not necessarily happen once in 30 years, it can happen more times or none at all. A 30-year flood has a 1-in-30 (30%) chance of happening in any one year.

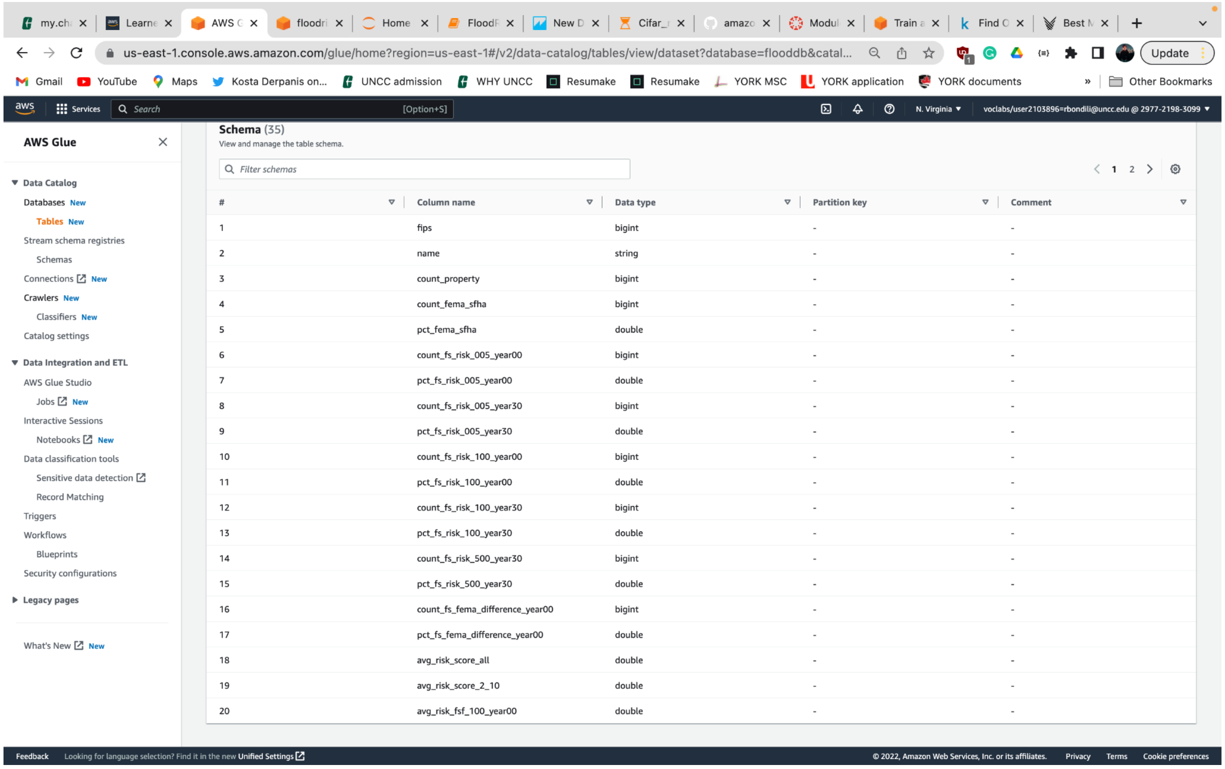
EXPLORATORY DATA ANALYSIS

We used the **AWS Sage Maker Notebook instance** for the exploratory data analysis and AWS Quick sight for data visualization which helps to gain insights on the data. Written code to perform null checks on rows and columns and see if there are any missing values from the four files in our Flood Risk Summary - Flood_risk_zipcodes; Flood_risk_cd; Flood_risk_county; Flood_risk_sate.

Data Understanding

To understand the schema of our data source, we have used **AWS Glue**. AWS Glue discovers our schema and defines the data types for the data. After analyzing the data, it stores the data in the data source. The Crawler feature is used to find the dataset schema. After seeing the schema and data types defined using Glue, AWS Athena is used for querying the data and finding the insights from the dataset.

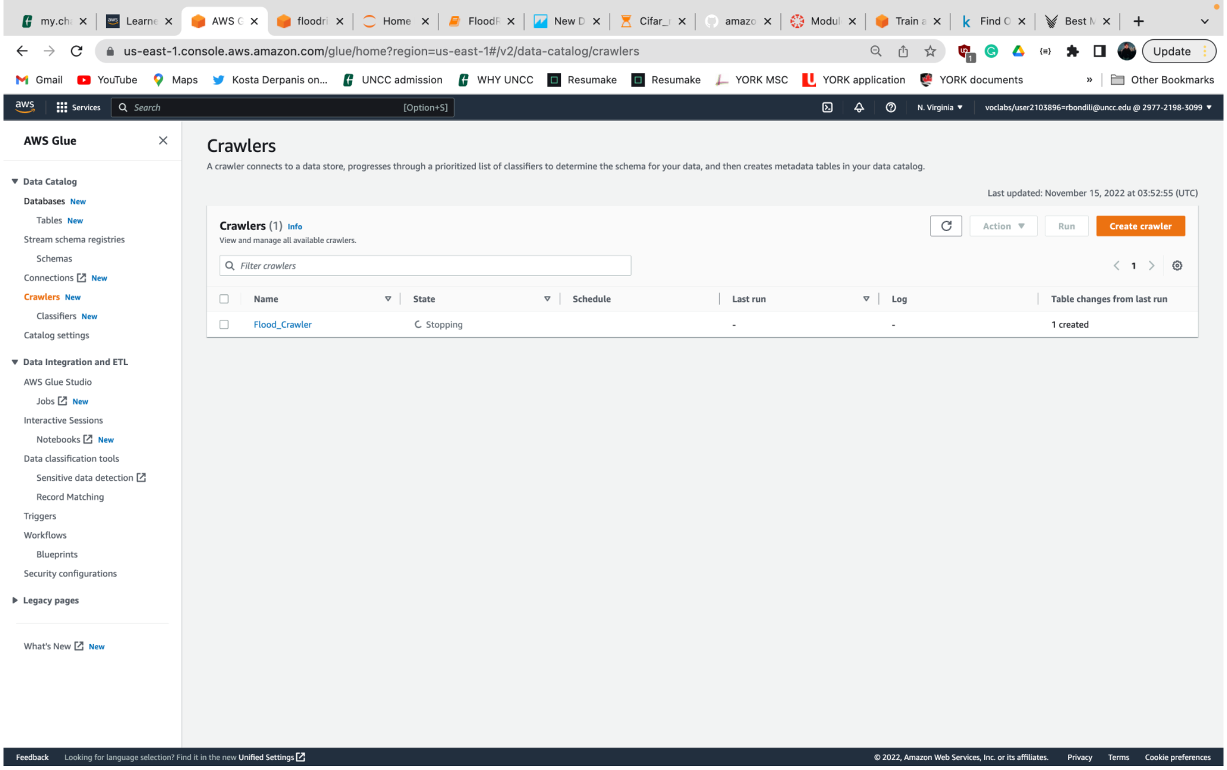
AWS GLUE



Schema (35)

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	fips	bigint	-	-
2	name	string	-	-
3	count_property	bigint	-	-
4	count_fema_sfha	bigint	-	-
5	pct_fema_sfha	double	-	-
6	count_fs_risk_005_year00	bigint	-	-
7	pct_fs_risk_005_year00	double	-	-
8	count_fs_risk_005_year30	bigint	-	-
9	pct_fs_risk_005_year30	double	-	-
10	count_fs_risk_100_year00	bigint	-	-
11	pct_fs_risk_100_year00	double	-	-
12	count_fs_risk_100_year30	bigint	-	-
13	pct_fs_risk_100_year30	double	-	-
14	count_fs_risk_500_year30	bigint	-	-
15	pct_fs_risk_500_year30	double	-	-
16	count_fs_fema_difference_year00	bigint	-	-
17	pct_fs_fema_difference_year00	double	-	-
18	avg_risk_score_all	double	-	-
19	avg_risk_score_2_10	double	-	-
20	avg_risk_fsf_100_year00	double	-	-



us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

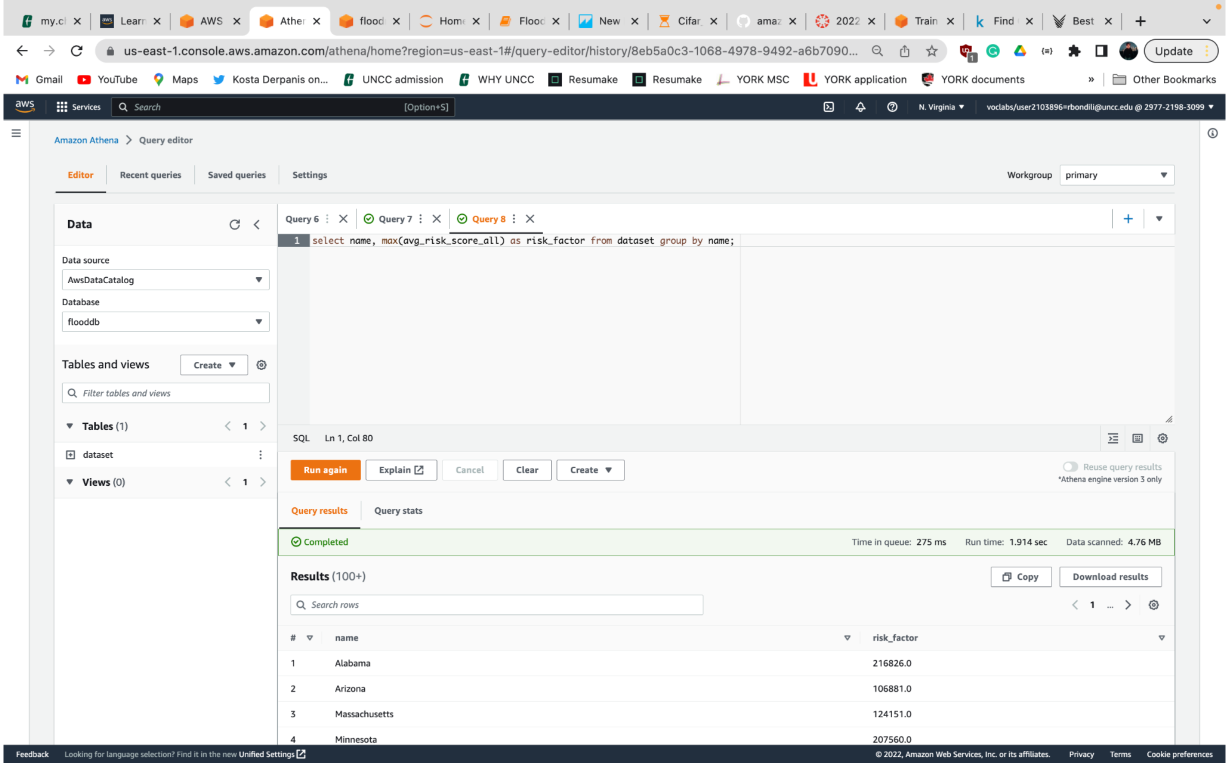
Last updated: November 15, 2022 at 03:52:55 (UTC)

Crawlers (1) Info

View and manage all available crawlers.

Name	State	Schedule	Last run	Log	Table changes from last run
Flood_Crawler	Stopping		-	-	1 created

AWS ATHENA



us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/8eb5a0c3-1068-4978-9492-a6b7090...

Amazon Athena > Query editor

Workgroup: primary

Data source: AwsDataCatalog

Database: flooddb

Tables (1): dataset

Views (0)

Query 8

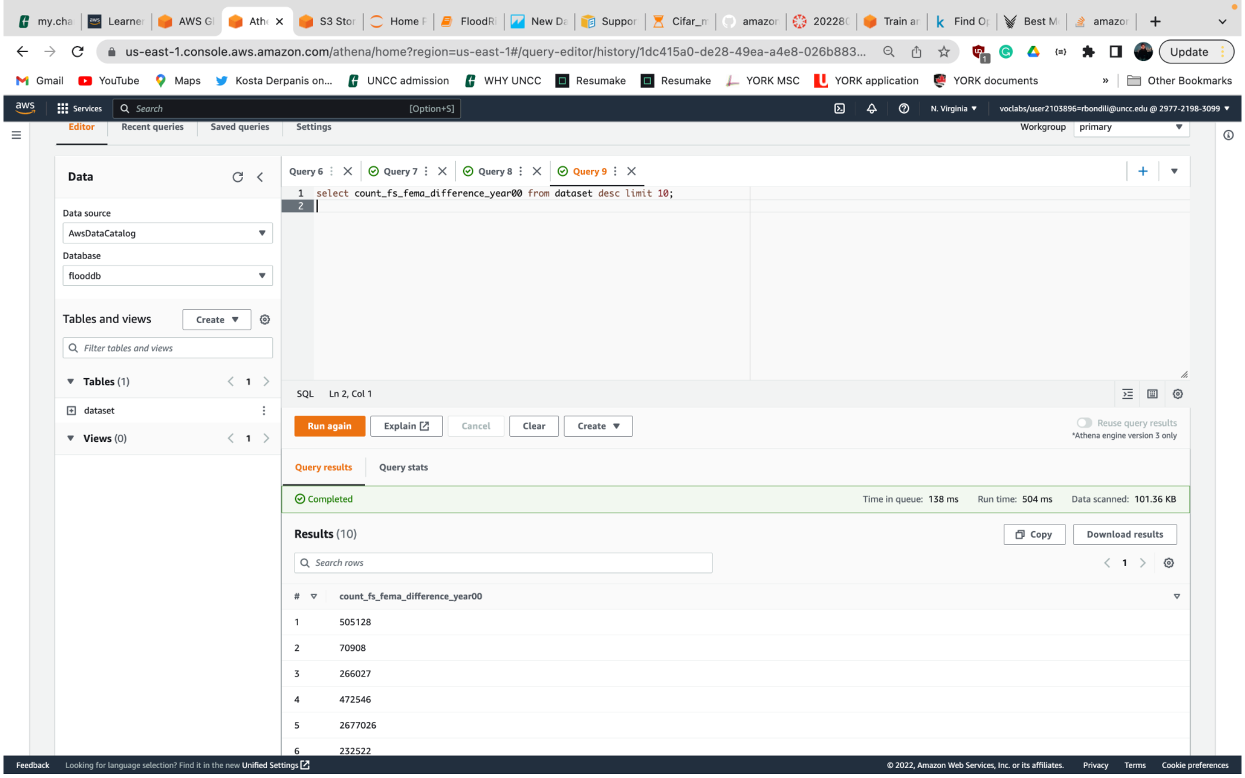
```
select name, max(avg_risk_score_all) as risk_factor from dataset group by name;
```

SQL Ln 1, Col 80

Completed Time in queue: 275 ms Run time: 1.914 sec Data scanned: 4.76 MB

Results (100+)

#	name	risk_factor
1	Alabama	216826.0
2	Arizona	106881.0
3	Massachusetts	124151.0
4	Minnesota	207560.0



us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/1dc415a0-de28-49ea-a4e8-026b883...

Data source: AwsDataCatalog

Database: flooddb

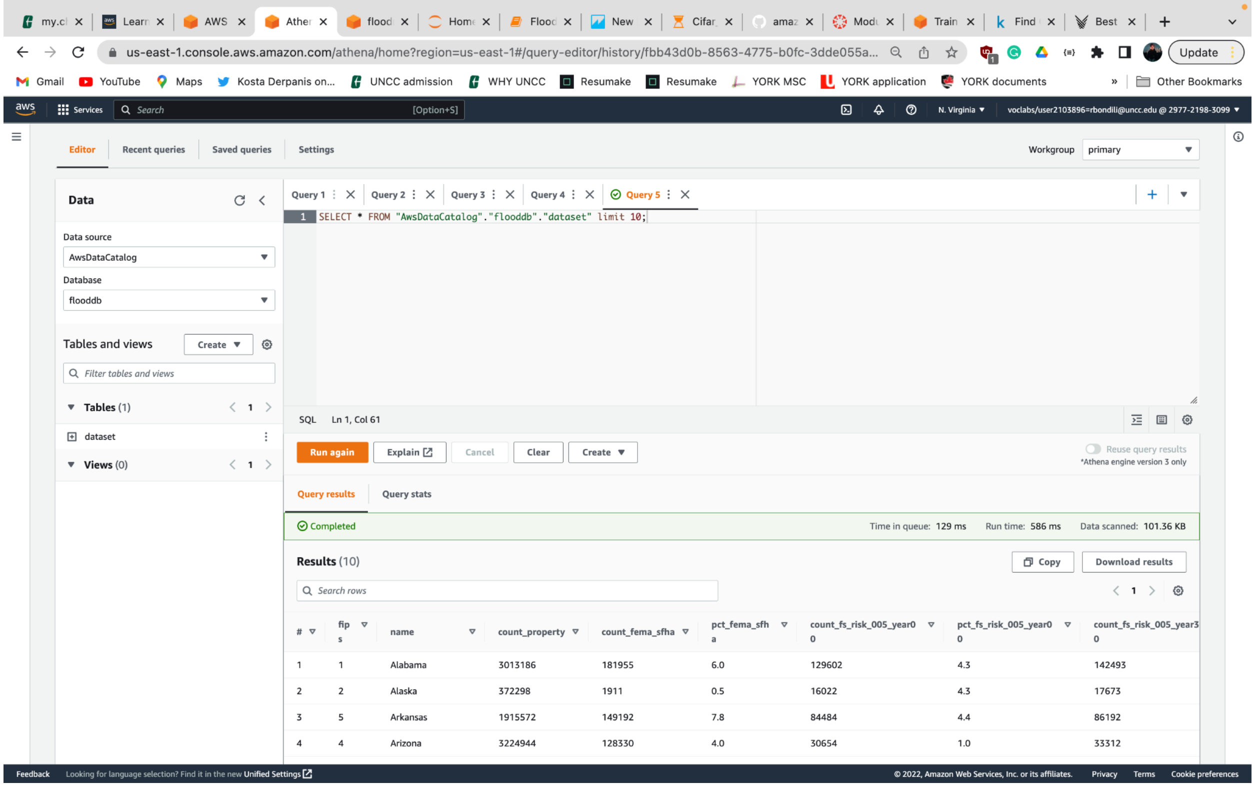
```
select count_fs_fema_difference_year00 from dataset desc limit 10;
```

SQL Ln 2, Col 1

Completed Time in queue: 138 ms Run time: 504 ms Data scanned: 101.36 KB

Results (10)

#	count_fs_fema_difference_year00
1	505128
2	70908
3	266027
4	472546
5	2677026
6	232522



us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/fbb43d0b-8563-4775-b0fc-3dde055a...

Data source: AwsDataCatalog

Database: flooddb

```
SELECT * FROM "AwsDataCatalog"."flooddb"."dataset" limit 10;
```

SQL Ln 1, Col 61

Completed Time in queue: 129 ms Run time: 586 ms Data scanned: 101.36 KB

Results (10)

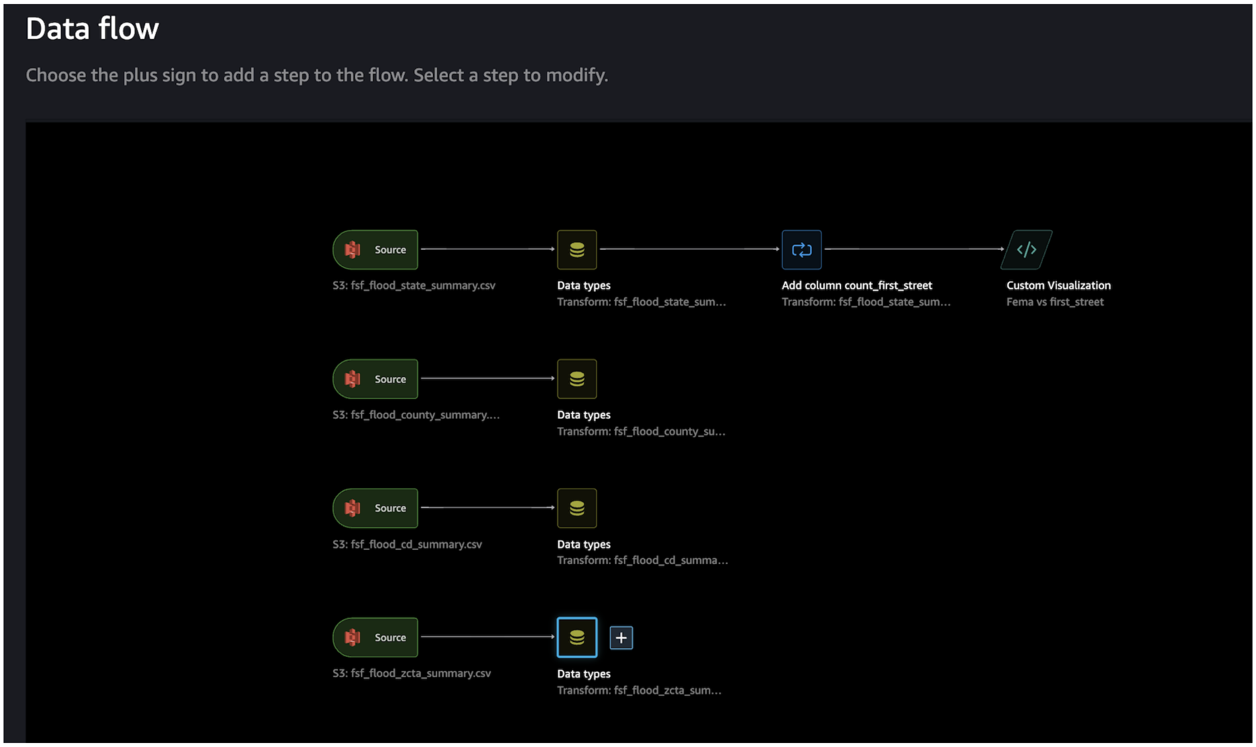
#	fips	name	count_property	count_fema_sfha	pct_fema_sfha	count_fs_risk_005_year00	pct_fs_risk_005_year00	count_fs_risk_005_year30
1	1	Alabama	3013186	181955	6.0	129602	4.3	142493
2	2	Alaska	372298	1911	0.5	16022	4.3	17673
3	5	Arkansas	1915572	149192	7.8	84484	4.4	86192
4	4	Arizona	3224944	128330	4.0	30654	1.0	33312

Data Preparation

For data modeling, we need to prepare the data and segregate it into three categories to set different levels of flood risks. We have used file - fsf_flood_zcta_summary to classify data. We have used AWS Sage maker Notebook Instance and using pandas, we have written the code to organize and group the data as required for the following data modeling steps. We have combined four columns- count_floodfactor1, count_floodfactor2, count_floodfactor3, and count_floodfactor4 to form a low-risk factor combined value. We combined count_floodfactor5, count_floodfactor6, and count_floodfactor7 cues data to derive a medium risk factor. Finally, for the high-risk factor value, we have used count_floodfactor8, count_floodfactor9, and count_floodfactor10.

Please find the data preparation and exploratory data analysis notebook in the repository.

We also utilized **Sagemaker studio Data Wrangler** to create a data flow by importing data from S3 bucket and performing data transformations, and visualizations.



Data flow

Choose the plus sign to add a step to the flow. Select a step to modify.

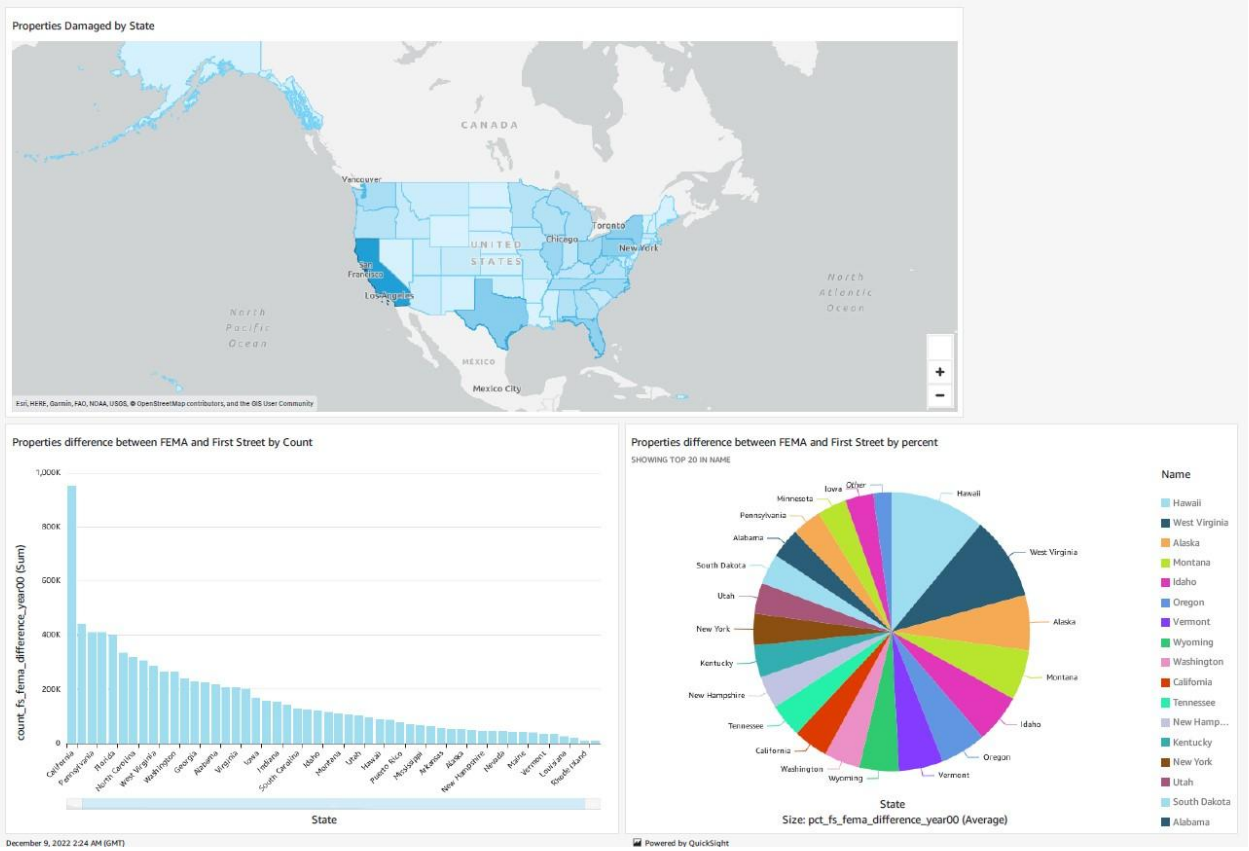
Source
S3: fsf_flood_state_summary.csv
Data types
Transform: fsf_flood_state_sum...
Add column count_first_street
Transform: fsf_flood_state_sum...
Custom Visualization
Fema vs first_street
Source
S3: fsf_flood_county_summary....
Data types
Transform: fsf_flood_county_su...
Source
S3: fsf_flood_cd_summary.csv
Data types
Transform: fsf_flood_cd_summa...
Source
S3: fsf_flood_zcta_summary.csv
+
Data types
Transform: fsf_flood_zcta_sum...

DATA VISUALIZATION

As a part of data visualization, we have utilized **AWS Quick Sight** service for creating a dashboard.

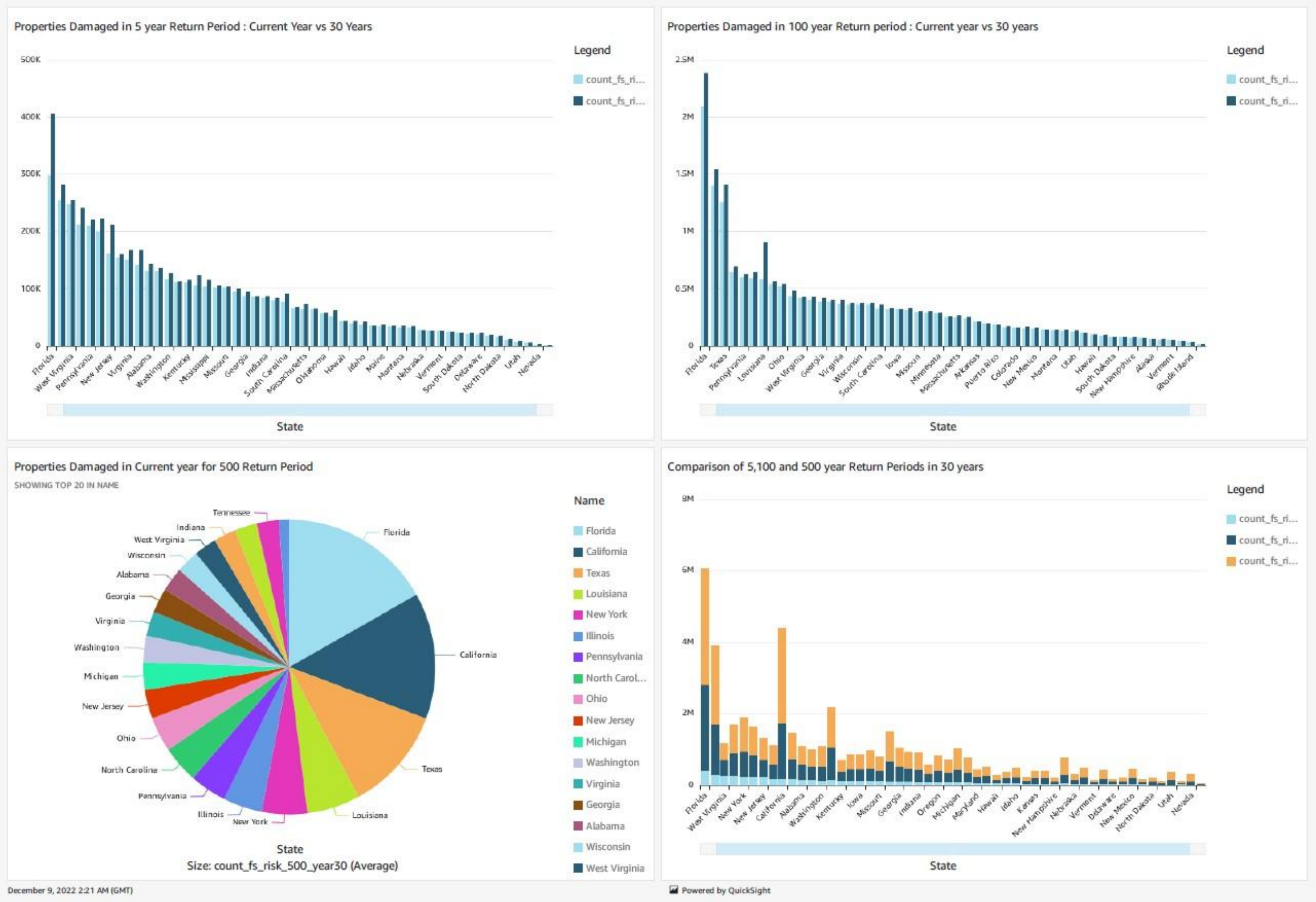
Properties damaged by state:

Here we use a map of America to visualize the number of properties damaged. This is indicated by the color tone of the state. A Histogram shows the Properties difference between FEMA and First Street by Count. We also have a Pie-chart that shows Properties difference between FEMA and First Street by Percentage.



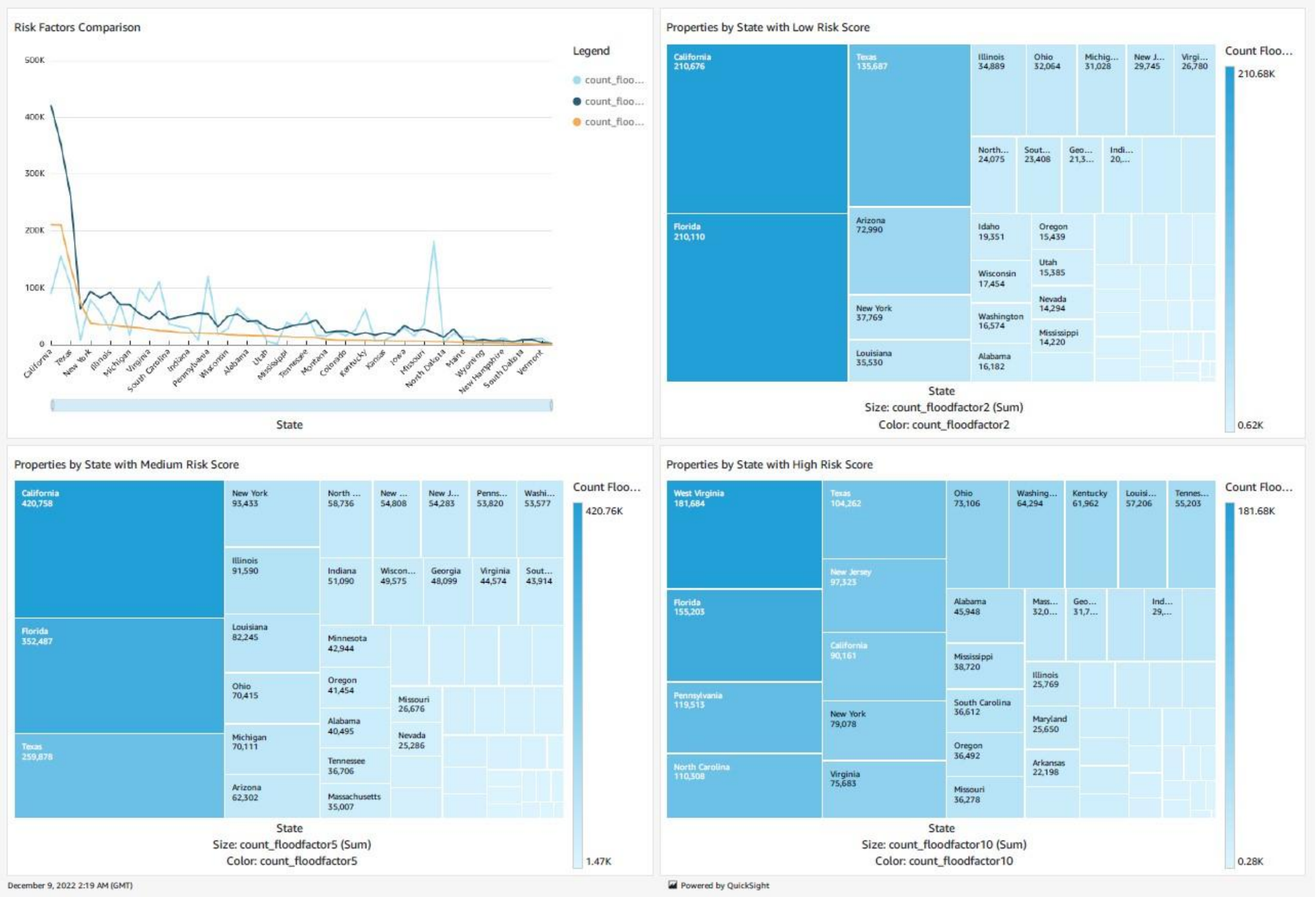
Return period comparison for 30 years:

In the second visualization, we predicted the occurrence of 5,100 and 500 return period flood possibilities in the next 30 years for all the states and counties. Here we make use of Bar charts and a pie chart to visualize the Properties Damaged in 5 years, 100 years, and 500 years Return Periods for the Current Year vs 30 Years. We also use a stacked bar chart to compare all return periods for 30 years.



Risk factor comparison:

This dashboard compares one risk factor each from high, medium, and low ranges. We have a line chart with risk factors 2, 5, and 10. There are also three treemaps for each of these separately.



DELIVERABLE - 3

For Deliverable 3, we explore various ML modeling through **AWS Sagemaker Canvas** and also **AWS Sagemaker Notebook Instance**. The data utilized is the zipcode data

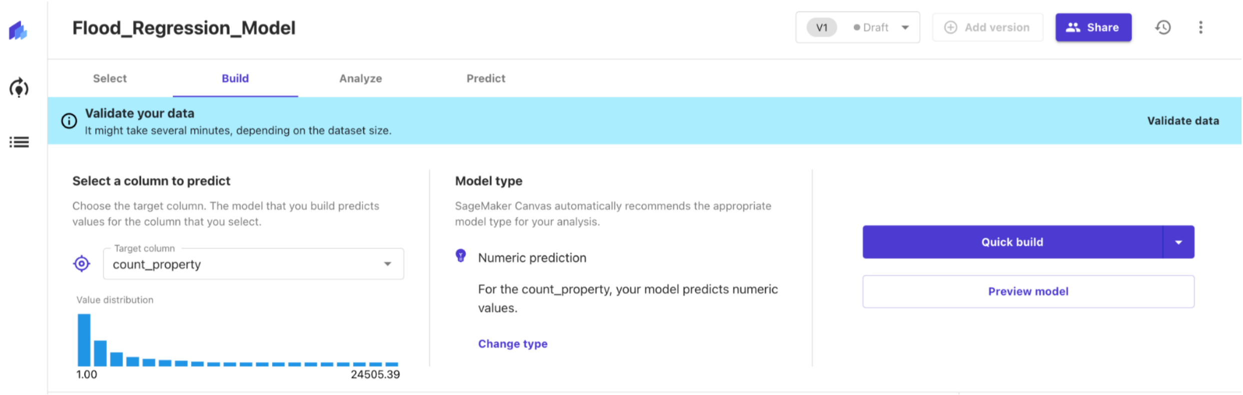
ANALYTICS AND MACHINE LEARNING

We developed a regression model to predict the number of properties destroyed by utilizing other data columns as features. We make use of the AWS Sagemaker canvas service to design a suitable regression model to predict the number of properties damaged.

We begin the canvas by selecting the type of problem, which would be regression in our case.

- Next, we are required to connect the data source which would be our zip code data file.
- Before modeling we are required to validate our data which would remove nulls and if required convert categorical variables into numeric vectors and also provide statistical

analysis on various columns. We next select our target column which is to be predicted which in our case is the **count property.**



Validate your data

It might take several minutes, depending on the dataset size.

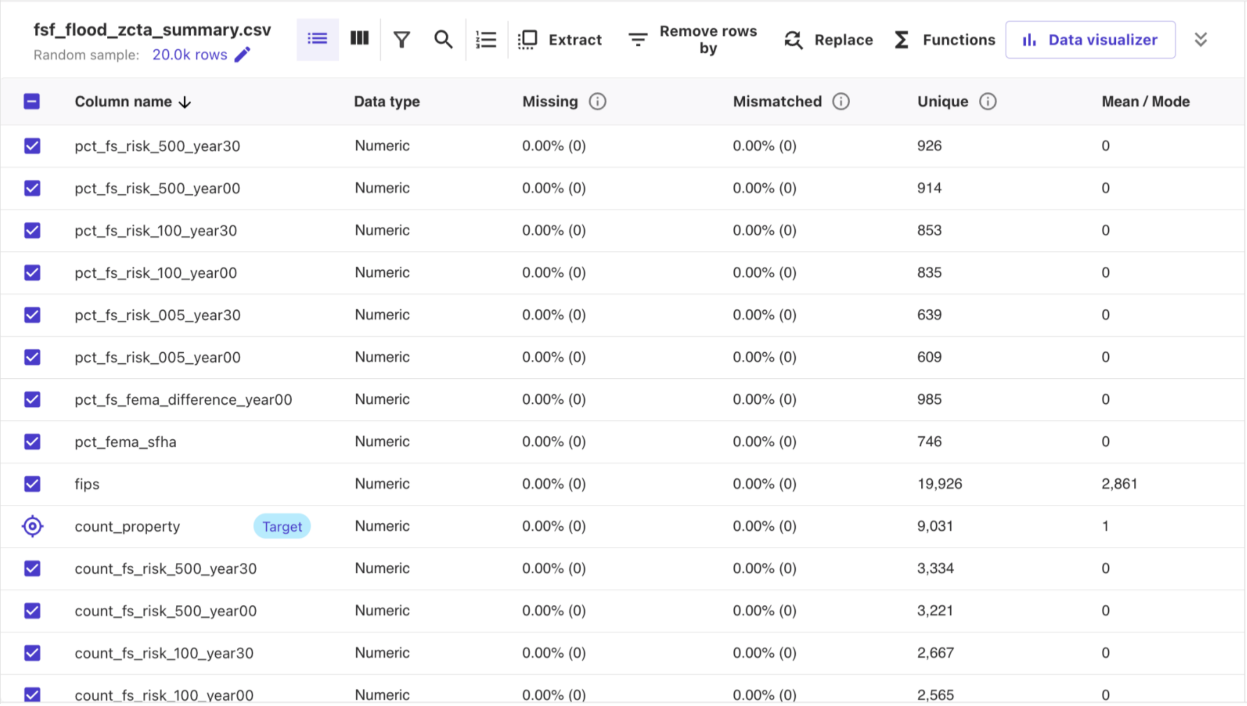
Select a column to predict

Choose the target column. The model that you build predicts values for the column that you select.

Model type

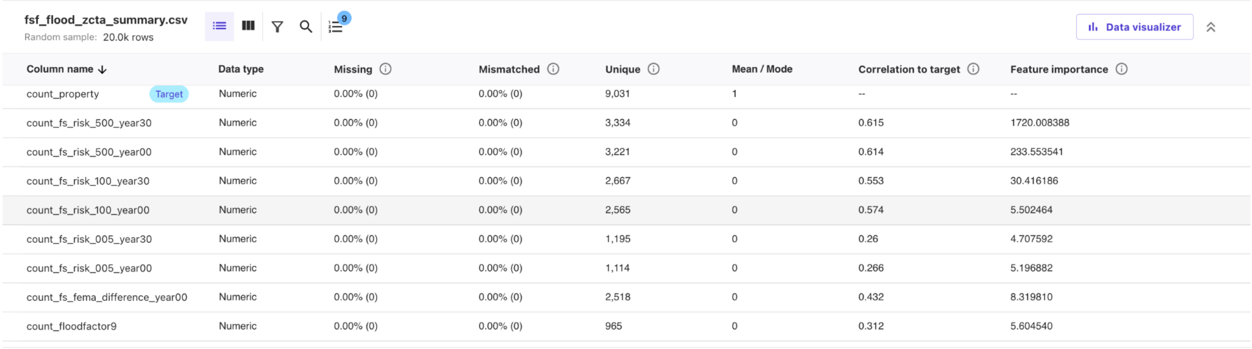
SageMaker Canvas automatically recommends the appropriate model type for your analysis.

For the count_property, your model predicts numeric values.



	Column name ↓	Data type	Missing	Mismatched	Unique	Mean / Mode
☑	pct_fs_risk_500_year30	Numeric	0.00% (0)	0.00% (0)	926	0
☑	pct_fs_risk_500_year00	Numeric	0.00% (0)	0.00% (0)	914	0
☑	pct_fs_risk_100_year30	Numeric	0.00% (0)	0.00% (0)	853	0
☑	pct_fs_risk_100_year00	Numeric	0.00% (0)	0.00% (0)	835	0
☑	pct_fs_risk_005_year30	Numeric	0.00% (0)	0.00% (0)	639	0
☑	pct_fs_risk_005_year00	Numeric	0.00% (0)	0.00% (0)	609	0
☑	pct_fs_fema_difference_year00	Numeric	0.00% (0)	0.00% (0)	985	0
☑	pct_fema_sfha	Numeric	0.00% (0)	0.00% (0)	746	0
☑	fips	Numeric	0.00% (0)	0.00% (0)	19,926	2,861
⌖	count_property Target	Numeric	0.00% (0)	0.00% (0)	9,031	1
☑	count_fs_risk_500_year30	Numeric	0.00% (0)	0.00% (0)	3,334	0
☑	count_fs_risk_500_year00	Numeric	0.00% (0)	0.00% (0)	3,221	0
☑	count_fs_risk_100_year30	Numeric	0.00% (0)	0.00% (0)	2,667	0
☑	count_fs_risk_100_year00	Numeric	0.00% (0)	0.00% (0)	2,565	0

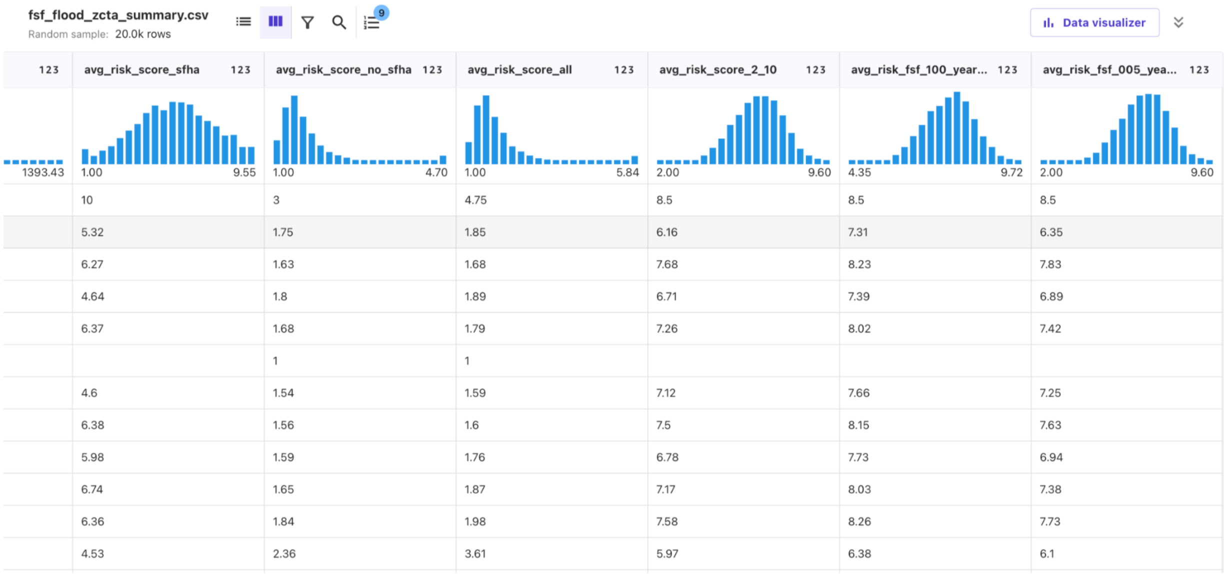
- In the next step we are required to select columns that are required features for the regression model.



fsf_flood_zcta_summary.csv

Random sample: 20.0k rows

Column name ↓	Data type	Missing	Mismatched	Unique	Mean / Mode	Correlation to target	Feature importance
count_property Target	Numeric	0.00% (0)	0.00% (0)	9,031	1	--	--
count_fs_risk_500_year30	Numeric	0.00% (0)	0.00% (0)	3,334	0	0.615	1720.008388
count_fs_risk_500_year00	Numeric	0.00% (0)	0.00% (0)	3,221	0	0.614	233.553541
count_fs_risk_100_year30	Numeric	0.00% (0)	0.00% (0)	2,667	0	0.553	30.416186
count_fs_risk_100_year00	Numeric	0.00% (0)	0.00% (0)	2,565	0	0.574	5.502464
count_fs_risk_005_year30	Numeric	0.00% (0)	0.00% (0)	1,195	0	0.26	4.707592
count_fs_risk_005_year00	Numeric	0.00% (0)	0.00% (0)	1,114	0	0.266	5.196882
count_fs_fema_difference_year00	Numeric	0.00% (0)	0.00% (0)	2,518	0	0.432	8.319810
count_floodfactor9	Numeric	0.00% (0)	0.00% (0)	965	0	0.312	5.604540



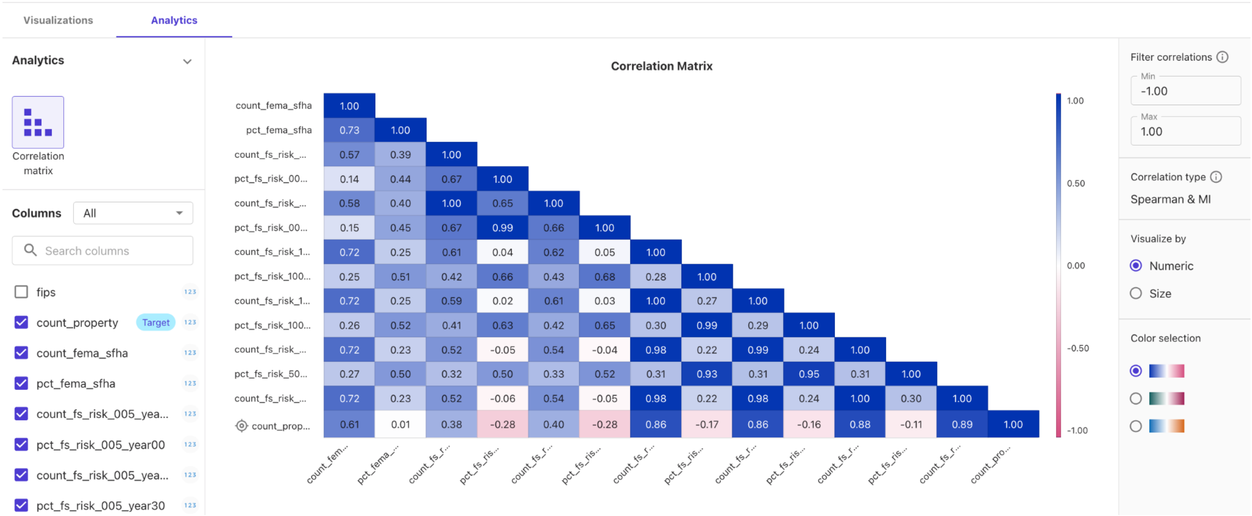
fsf_flood_zcta_summary.csv

Random sample: 20.0k rows

123	avg_risk_score_sfha 123	avg_risk_score_no_sfha 123	avg_risk_score_all 123	avg_risk_score_2_10 123	avg_risk_fsf_100_year... 123	avg_risk_fsf_005_yea... 123
1393.43	1.00 9.55	1.00 4.70	1.00 5.84	2.00 9.60	4.35 9.72	2.00 9.60
	10	3	4.75	8.5	8.5	8.5
	5.32	1.75	1.85	6.16	7.31	6.35
	6.27	1.63	1.68	7.68	8.23	7.83
	4.64	1.8	1.89	6.71	7.39	6.89
	6.37	1.68	1.79	7.26	8.02	7.42
		1	1			
	4.6	1.54	1.59	7.12	7.66	7.25
	6.38	1.56	1.6	7.5	8.15	7.63
	5.98	1.59	1.76	6.78	7.73	6.94
	6.74	1.65	1.87	7.17	8.03	7.38
	6.36	1.84	1.98	7.58	8.26	7.73
	4.53	2.36	3.61	5.97	6.38	6.1

The tool provides various visualizations to understand the features that help with data modeling.

Below heat map represents the correlations among the various features:



Analytics

Columns: All

Search columns

- □ fips
- ☑ count_property Target
- ☑ count_fema_sfha
- ☑ pct_fema_sfha
- ☑ count_fs_risk_005_yea...
- ☑ pct_fs_risk_005_year00
- ☑ count_fs_risk_005_yea...
- ☑ pct_fs_risk_005_year30

Correlation Matrix

	count_fem...	pct_fema_...	count_fs_r...	pct_fs_ris...	count_fs_r...	pct_fs_ris...	count_fs_r...	pct_fs_ris...	count_fs_r...	pct_fs_ris...	count_fs_r...	pct_fs_ris...	count_fs_r...	count_pro...
count_fema_sfha	1.00													
pct_fema_sfha	0.73	1.00												
count_fs_risk_...	0.57	0.39	1.00											
pct_fs_risk_00...	0.14	0.44	0.67	1.00										
count_fs_risk_...	0.58	0.40	1.00	0.65	1.00									
pct_fs_risk_00...	0.15	0.45	0.67	0.99	0.66	1.00								
count_fs_risk_1...	0.72	0.25	0.61	0.04	0.62	0.05	1.00							
pct_fs_risk_100...	0.25	0.51	0.42	0.66	0.43	0.68	0.28	1.00						
count_fs_risk_1...	0.72	0.25	0.59	0.02	0.61	0.03	1.00	0.27	1.00					
pct_fs_risk_100...	0.26	0.52	0.41	0.63	0.42	0.65	0.30	0.99	0.29	1.00				
count_fs_risk_...	0.72	0.23	0.52	-0.05	0.54	-0.04	0.98	0.22	0.99	0.24	1.00			
pct_fs_risk_50...	0.27	0.50	0.32	0.50	0.33	0.52	0.31	0.93	0.31	0.95	0.31	1.00		
count_fs_risk_...	0.72	0.23	0.52	-0.06	0.54	-0.05	0.98	0.22	0.98	0.24	1.00	0.30	1.00	
count_prop...	0.61	0.01	0.38	-0.28	0.40	-0.28	0.86	-0.17	0.86	-0.16	0.88	-0.11	0.89	1.00

Filter correlations
Min -1.00
Max 1.00

Correlation type
Spearman & MI

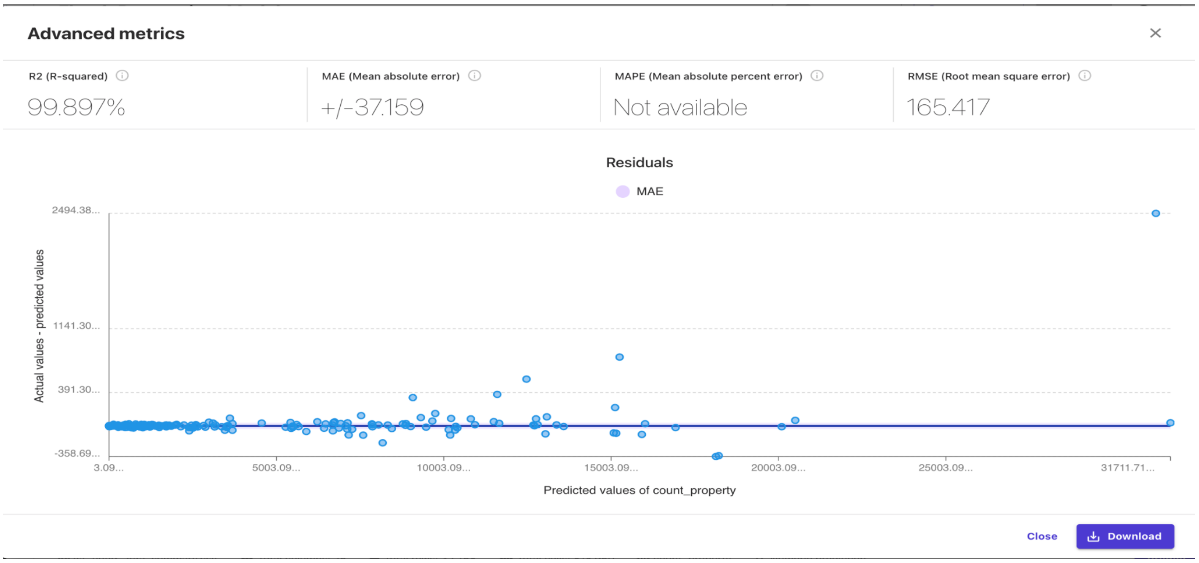
Visualize by
● Numeric
○ Size

Color selection
●
○
○

We then utilize the Quick build option, to generate a linear regression model which takes features as input data and fits the model to the data.

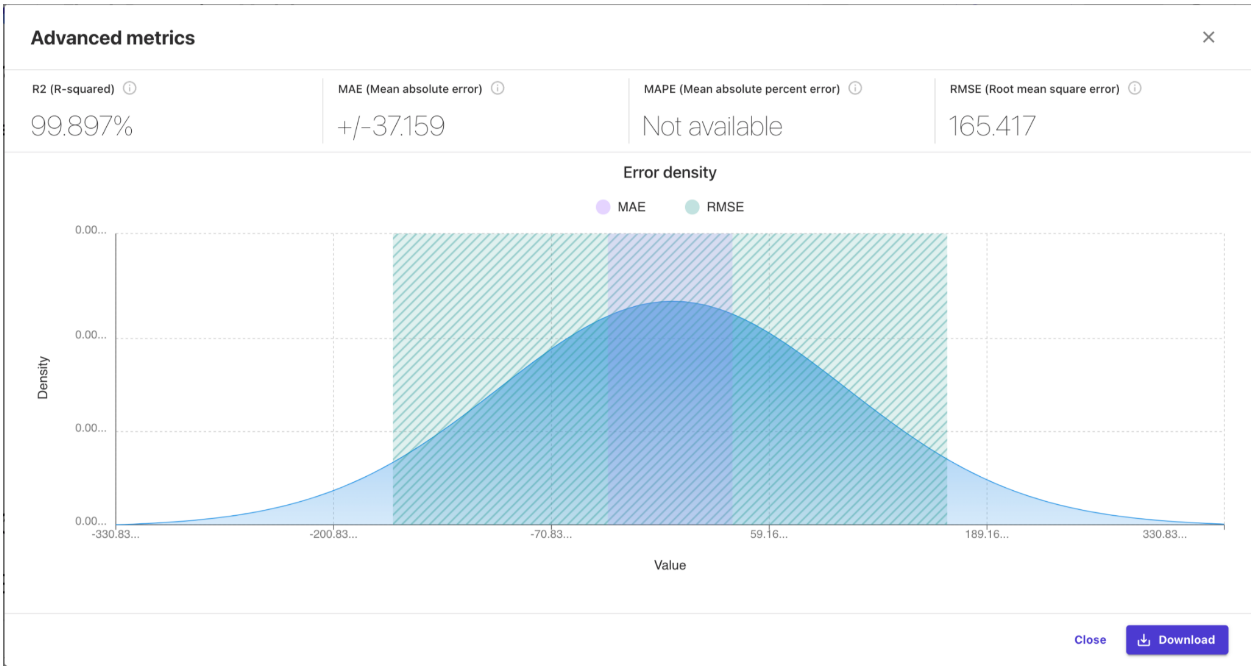
EVALUATION AND OPTIMIZATION

The model evaluation is done on various metrics such as MAE, MSE, RMSE, etc. Among the various metrics, the best results are selected for the model.



Advanced metrics

R2 (R-squared)	MAE (Mean absolute error)	MAPE (Mean absolute percent error)	RMSE (Root mean square error)
99.897%	+/-37.159	Not available	165.417



Advanced metrics

R2 (R-squared): 99.897%

MAE (Mean absolute error): +/-37.159

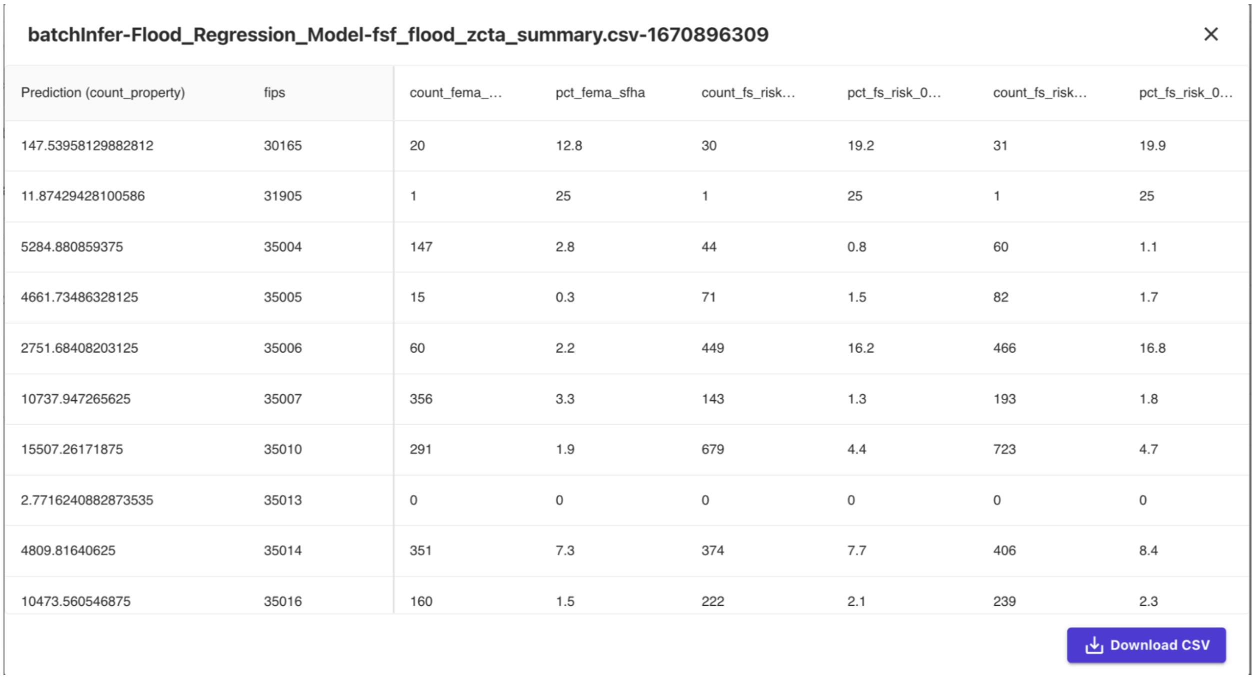
MAPE (Mean absolute percent error): Not available

RMSE (Root mean square error): 165.417

Predictions are made using the selected model and the test score is evaluated on the selected metric.

PREDICTIONS

Following are the model prediction results.



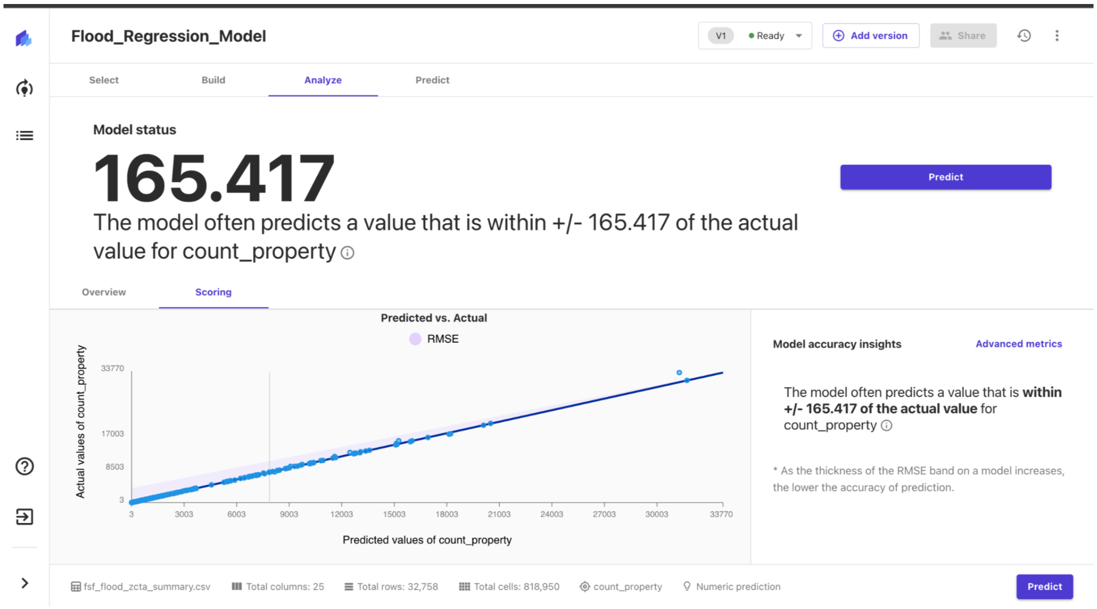
batchInfer-Flood_Regression_Model-fsf_flood_zcta_summary.csv-1670896309

Prediction (count_property)	fips	count_fema_...	pct_fema_sfha	count_fs_risk...	pct_fs_risk_0...	count_fs_risk...	pct_fs_risk_0...
147.53958129882812	30165	20	12.8	30	19.2	31	19.9
11.87429428100586	31905	1	25	1	25	1	25
5284.880859375	35004	147	2.8	44	0.8	60	1.1
4661.73486328125	35005	15	0.3	71	1.5	82	1.7
2751.68408203125	35006	60	2.2	449	16.2	466	16.8
10737.947265625	35007	356	3.3	143	1.3	193	1.8
15507.26171875	35010	291	1.9	679	4.4	723	4.7
2.7716240882873535	35013	0	0	0	0	0	0
4809.81640625	35014	351	7.3	374	7.7	406	8.4
10473.560546875	35016	160	1.5	222	2.1	239	2.3

The complete predictions file is available in the repository.

RESULTS

The model generated is a simple linear regression model and it is evaluated on RMSE. The value of RMSE for our model is 165.417



Model status

165.417

The model often predicts a value that is within +/- 165.417 of the actual value for count_property

The model often predicts a value that is **within +/- 165.417 of the actual value** for count_property

* As the thickness of the RMSE band on a model increases, the lower the accuracy of prediction.

fsf_flood_zcta_summary.csv | Total columns: 25 | Total rows: 32,758 | Total cells: 818,950 | count_property | Numeric prediction

FUTURE WORK AND COMMENTS

It is important to clean the data sample in the modeling process to ensure that the observations best represent the problem. The presence of outliers in a dataset can result in a poor fit and lower predictive modeling performance. In the current data we had some columns which were not necessary for prediction of data.

The steps we followed for cleaning as well as checking nulls in the data were

- Standardizing data
- Checking Nan or nulls in the data
- Validating the data
- Analyzing data quality

We had to remove the rows which were falling in the outlier area so that it does not mess the prediction. This helped us a lot for visualizations as well.
The four csv files used in our project were connected with each other with the primary key as fips.

Two new columns were necessary for generating the visualizations for the other organization because only one of the columns was given. This data can be used for visualizing the current state of the country. This visualized data can be updated regularly if the upstream of the data link is updated.

The current project can be used with weather updation services on flood data analysis based on previous data. The design of the AWS services can be used for other projects as well.