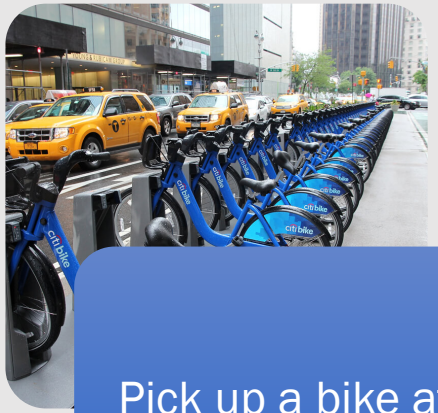
BAY AREA BIKE RECOMMENDATION APP

TEAM SPARKLE

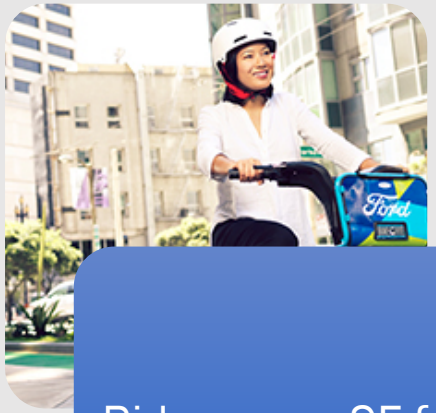
ANANT AGARWAL
ARPITA JENA
ASMITA VIKAS
DEENA LIZ JOHN

- INTRODUCTION
- RECOMMENDATION APP
- DATA DESCRIPTION
- IMPLEMENTATION IN SPARK
- MODEL BUILDING
- FUTURE SCOPE
- LESSONS LEARNED

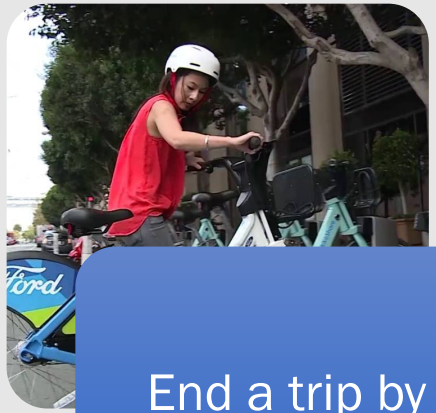
Introduction



Pick up a bike at any Ford GoBike station.



Ride across SF for quick trips!



End a trip by returning at any bike station.

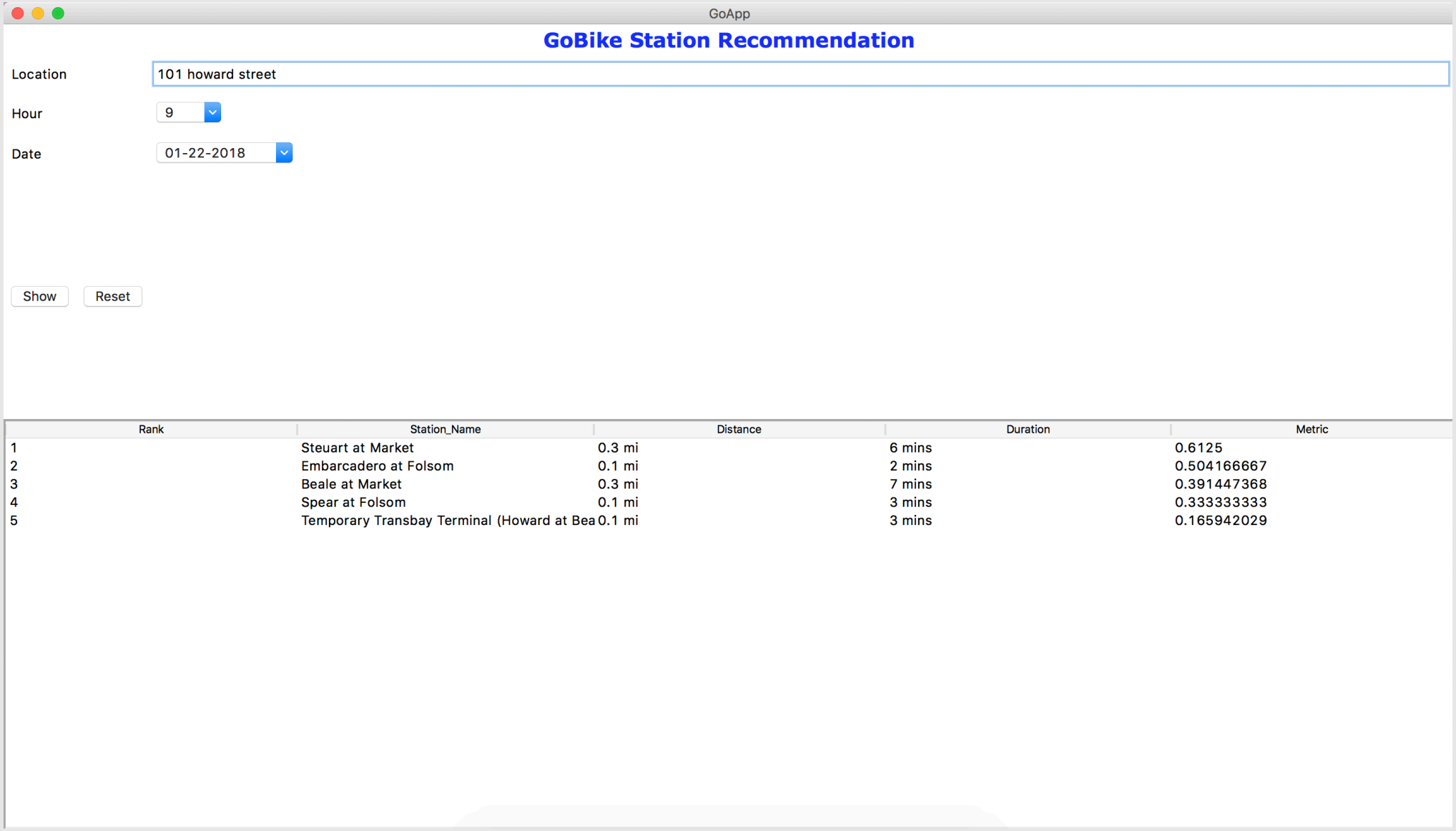
Business Problem

- Given the shortage of bikes, there is a need to optimize availability of bikes to users.
- With the existing Ford GoBike App, real-time bike availability information can be tracked.
- Need a tool to plan future rides.

Recommendation System

- System finds nearest station for future rides based on traffic and bike availability using Google Maps API.
- Users are provided a ranked list of these stations based on predicted bike availability and traffic.

DEMO


GoApp
GoBike Station Recommendation
Location
101 howard street
Hour
9
Date
01-22-2018
Show
Reset
Rank
Station_Name
Distance
Duration
Metric
1
Steuart at Market
0.3 mi
6 mins
0.6125
2
Embarcadero at Folsom
0.1 mi
2 mins
0.504166667
3
Beale at Market
0.3 mi
7 mins
0.391447368
4
Spear at Folsom
0.1 mi
3 mins
0.333333333
5
Temporary Transbay Terminal (Howard at Bea
0.1 mi
3 mins
0.165942029

Data Description

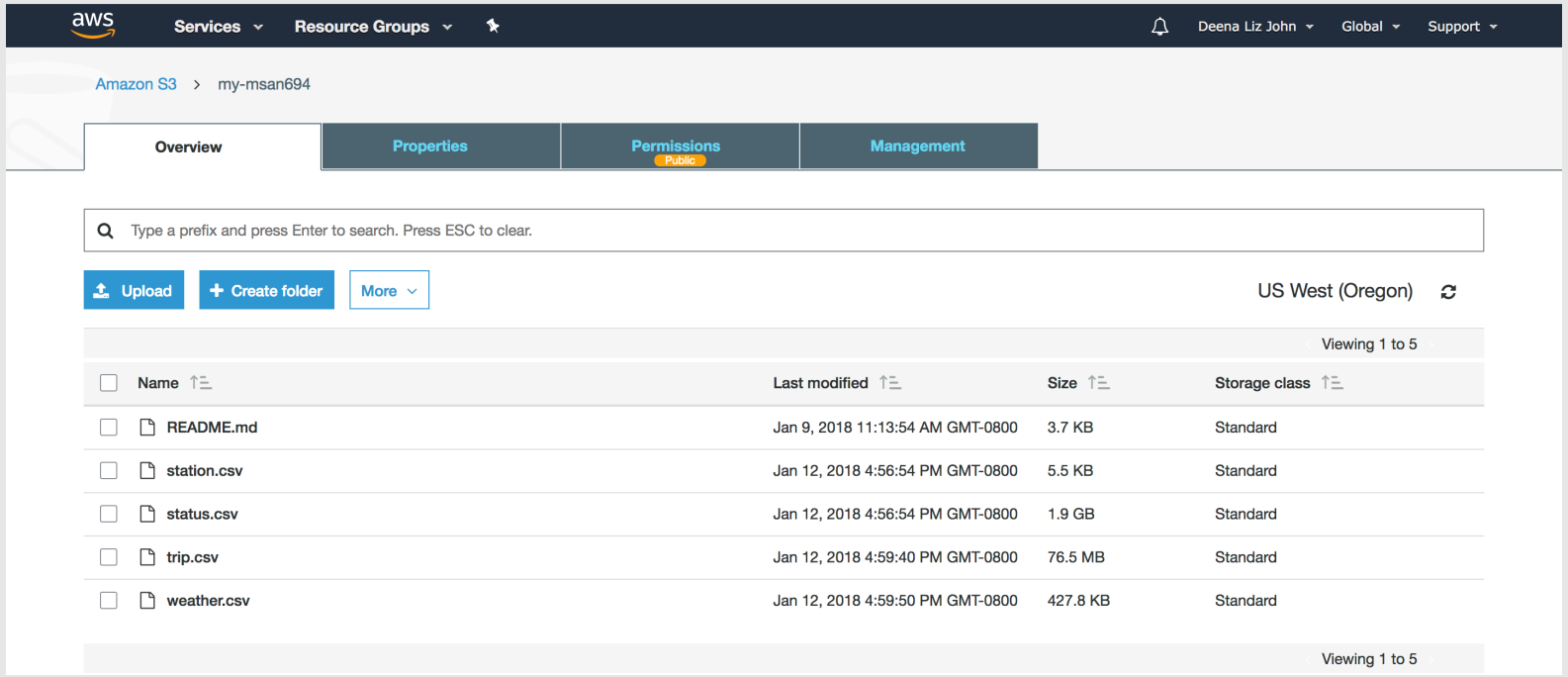
- 2.5 GB data from Ford GoBike data release with trip and station information.
- Data from August 2013 to August 2015.
- Nearly 71 million trips.
- Over 15 features of weather information was used as external data.

Pipeline

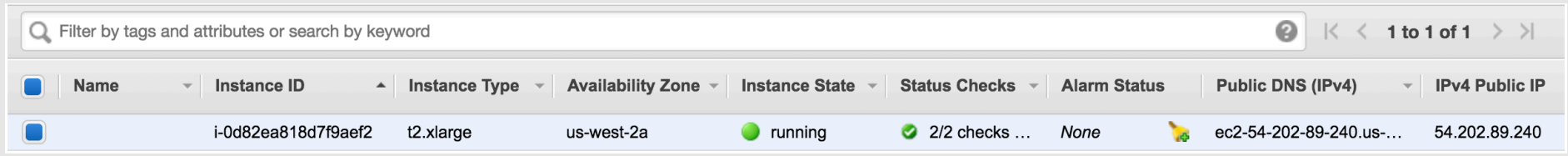

Data
amazon web services S3
mongoDB.
APACHE Spark
amazon web services EC2

Implementation in Spark

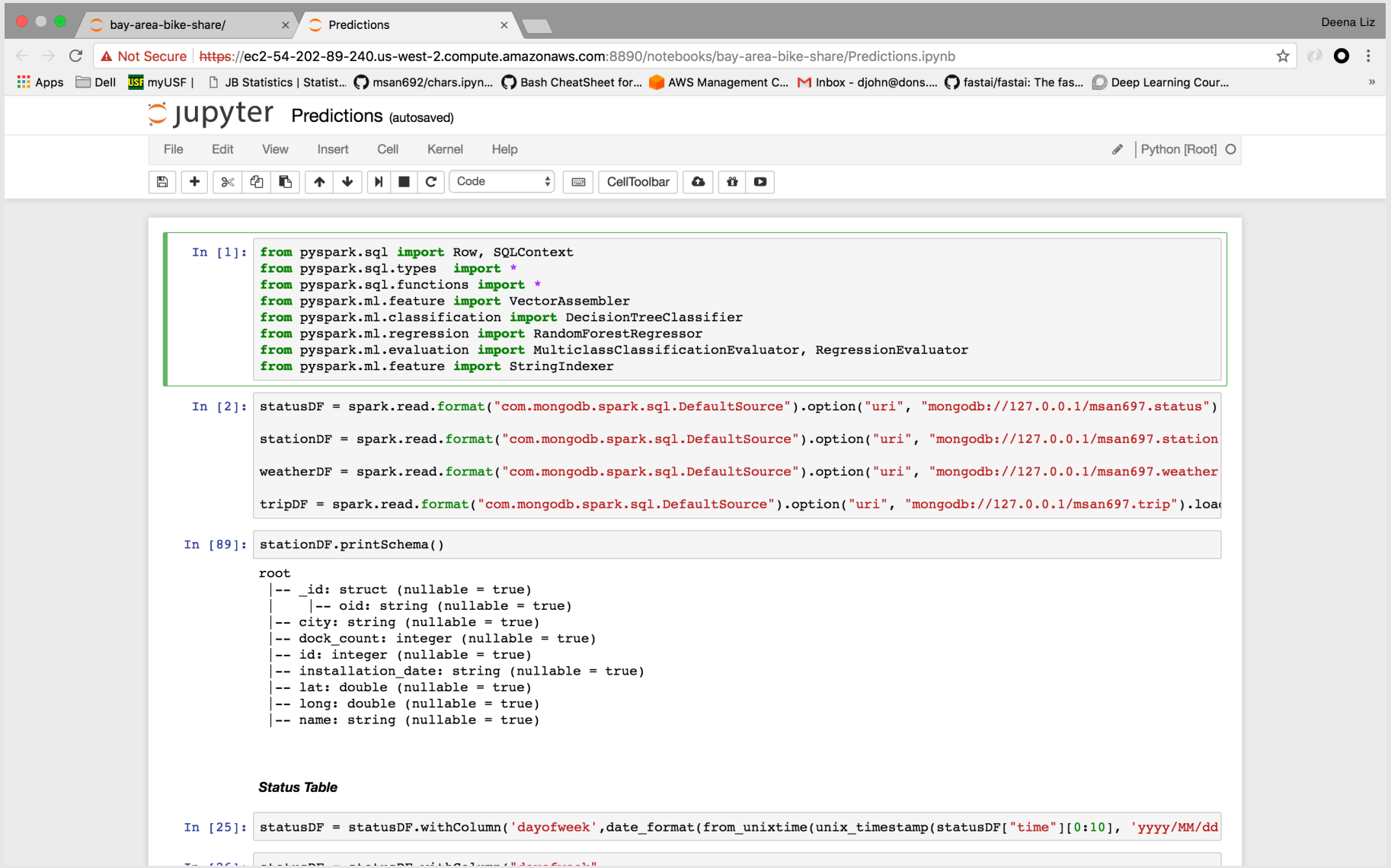
Data loaded in S3



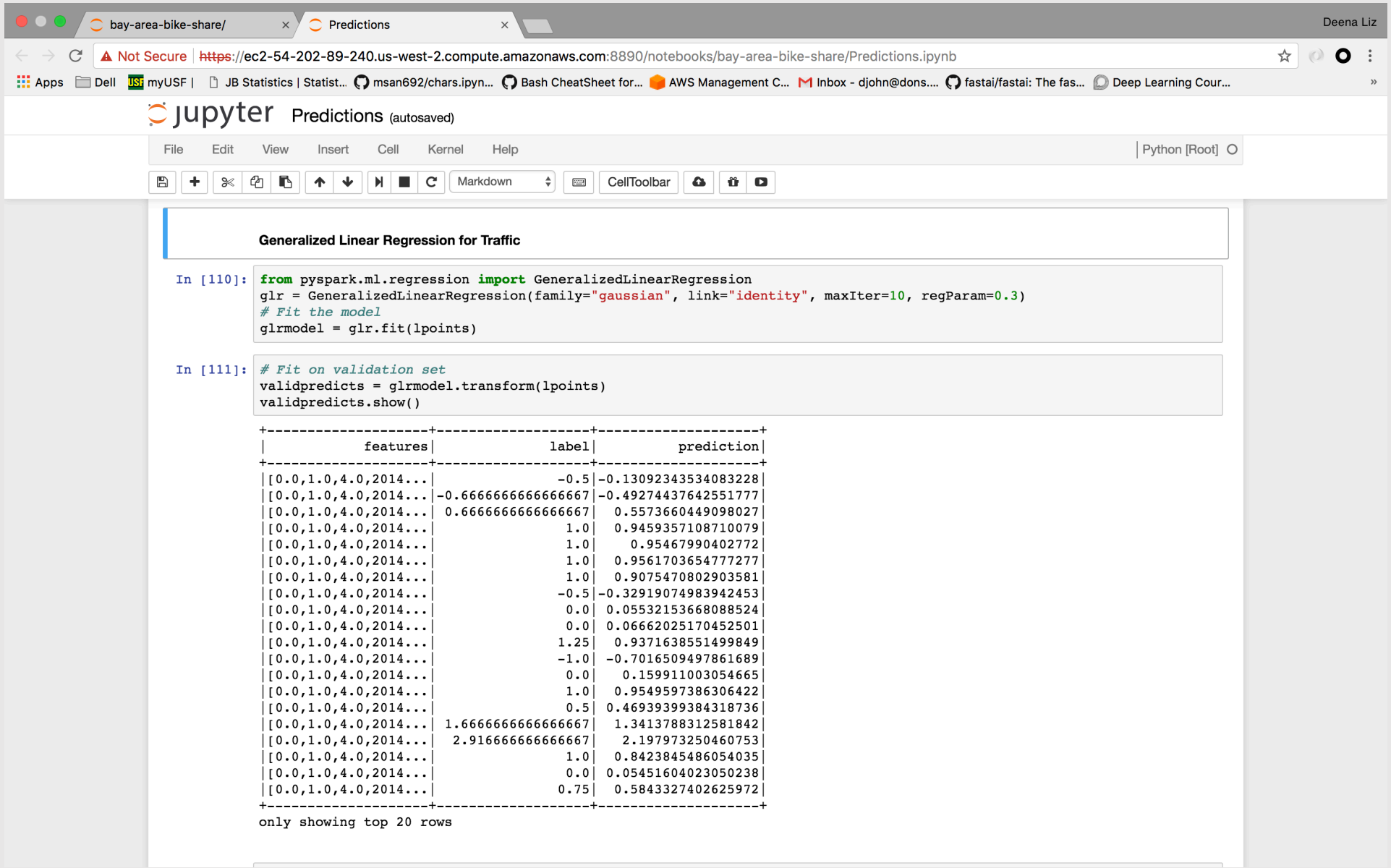
EC2 Instance



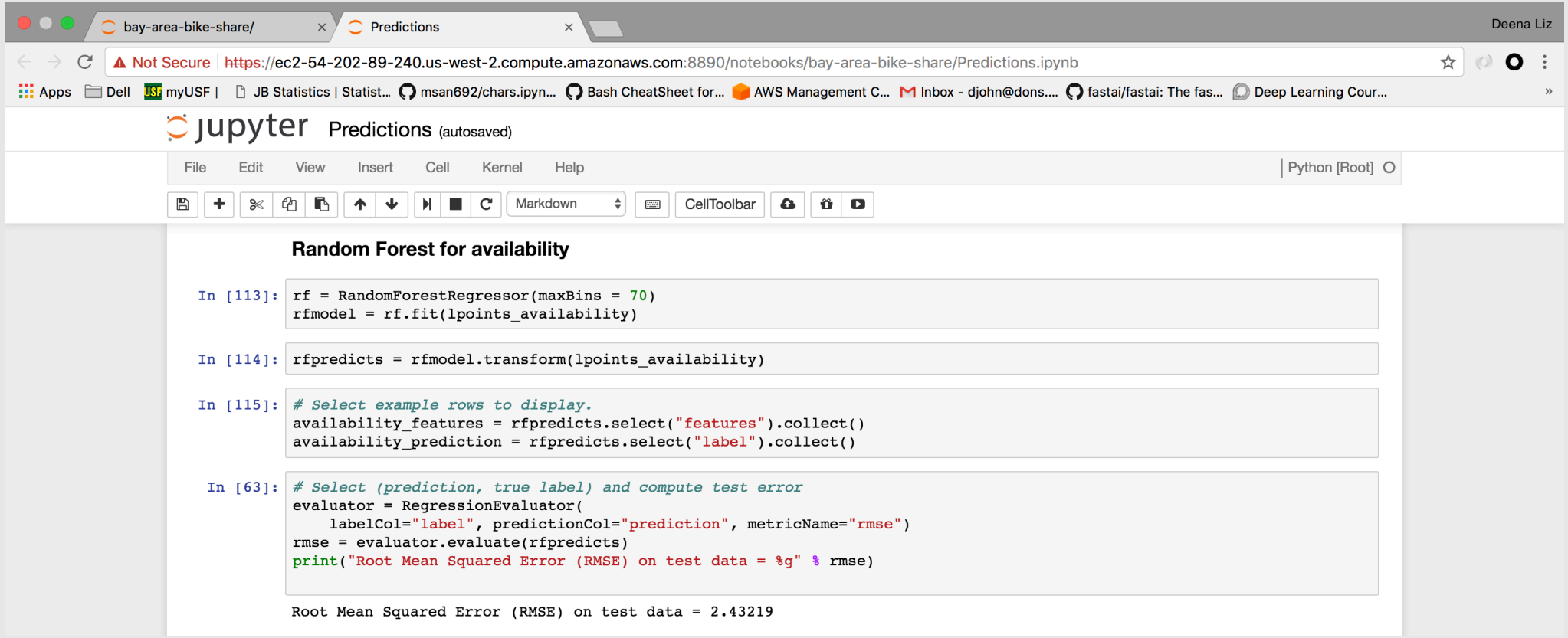
Querying on EC2



Implementing Generalized Linear Regression model to predict Traffic



Implementing Random Forest model to predict Availability



Future Scope

- Adding external datasets like SF crime data and topology data to improve model.
- Integrate real-time bike availability to update suggested future rides.
- Track location of user to suggest new stations.
- Recommend stations for bike drop-offs based on dock availability before even starting a trip.

Lessons Learned

- Cluster capacity and CPU memory should be in proportion with the volume of data.
- Caching the final tables after processing helped in quick modeling, avoiding reruns of the entire code.
- Writing SQL queries for feature engineering was easier by registering dataframes in the table catalog.
- Correct merge conflicts for Git repositories!

THANK YOU