Foundations of Data Management

INF- 551 Spring 2020

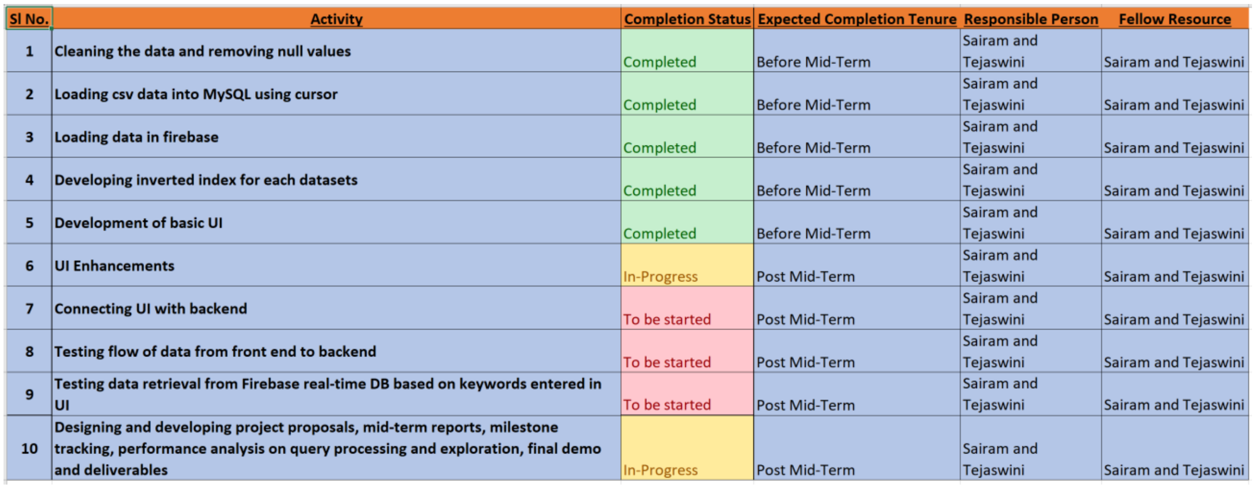
Keyword-Driven Exploration of Relational Data Using Firebase

Midterm Project Report

Sairam Kamal Raj Tejaswini Prakash Kulkarni

Contents

Current Status of Project



Sl No.	Activity	Completion Status	Expected Completion Tenure	Responsible Person	Fellow Resource
1	Cleaning the data and removing null values	Completed	Before Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
2	Loading csv data into MySQL using cursor	Completed	Before Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
3	Loading data in firebase	Completed	Before Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
4	Developing inverted index for each datasets	Completed	Before Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
5	Development of basic UI	Completed	Before Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
6	UI Enhancements	In-Progress	Post Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
7	Connecting UI with backend	To be started	Post Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
8	Testing flow of data from front end to backend	To be started	Post Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
9	Testing data retrieval from Firebase real-time DB based on keywords entered in UI	To be started	Post Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini
10	Designing and developing project proposals, mid-term reports, milestone tracking, performance analysis on query processing and exploration, final demo and deliverables	In-Progress	Post Mid-Term	Sairam and Tejaswini	Sairam and Tejaswini

About the datasets used:

As decided in the project proposal stage, we have been working on the following 2 datasets:

1. SF Bay Area Bike Share
2. Questions from Cross Validated Stack Exchange

Apart from the above mentioned data sets we also decided to analyze and use Novel Corona dataset which is an ongoing pandemic.

We have also loaded the “World” data set provided by the instructional faculty.

Steps carried out in order to achieve the above-mentioned status:

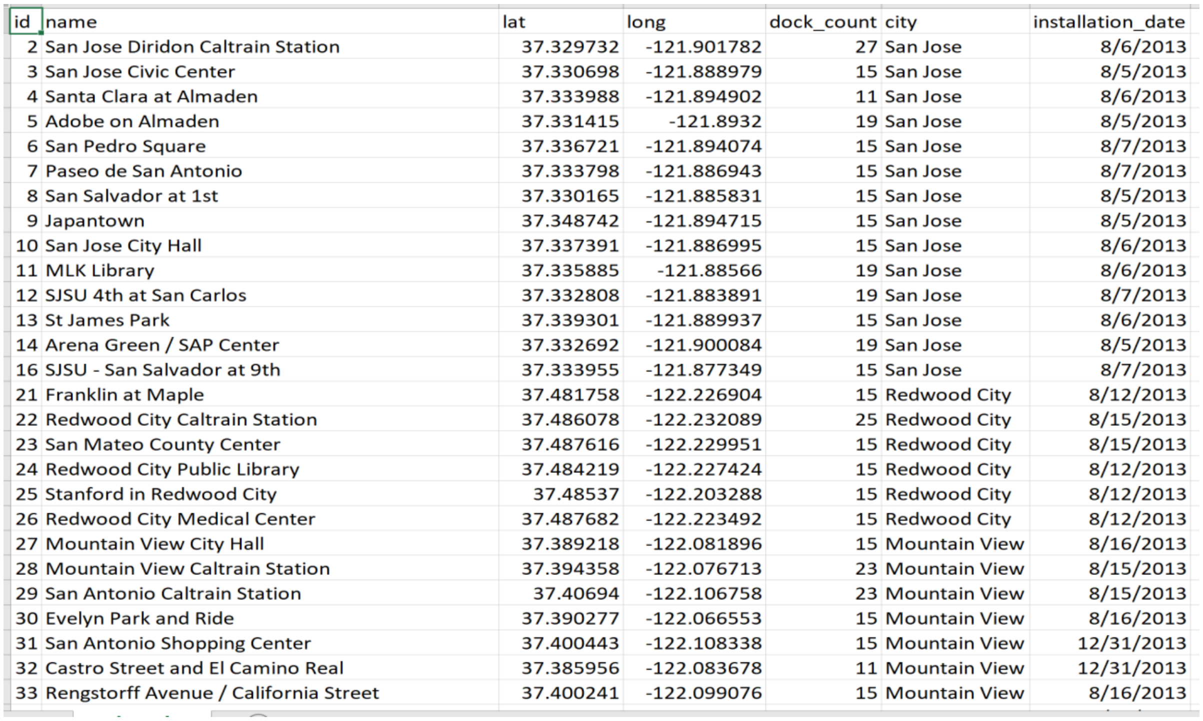
1. Loaded the data from kaggle onto a dataframe and cleaned the data.
2. Data cleansing process involved the following steps:
    i. Ensuring that the correct file format was used to read the files in order to overcome errors like file not been able to decode utf-8 and avoiding special characters which hinders loading data onto Firebase.
    ii. Formatting all the column names and renaming them incase if they are dirty: lowercase and removing of special characters (except underscore and white space).
    iii. Obtaining column datatypes for all the features and ensuring that the data follows the domain constraints.

    iv. Converting the format of date columns to YYYY-MM-DD, a format that is in compliant with MySQL database.
    v. Removing html specific tags such as <p>, </p> etc… that were present in stack overflow data.
    vi. Dropped duplicate rows in the data.
    vii. Cleansing of null values by replacing with 0’s or “Unavailable” based on the datatype of the column.
    viii. Testing of removal of data that hinders loading onto firebase.
3. Setting up of MySQL database in EC2 with root credentials and created the required tables and provided all privileges to user INF551.
4. Automated the process of loading the values in MySQL from the datasets using Python code.
5. Automated the process of extracting data from MySQL database using python using input parameters as database name and the name of the database to be written in firebase.
6. Created JSON data from the data which is present on MySQL using Python scripts.
7. Utilized the created JSON data to load on to the firebase.
8. Created inverted index for each dataset which included the following steps:
    i. Reading the text data from each of the datasets individually.
    ii. Cleansing the text data by converting to lowercase, removing the special characters and splitting the data based on space.
    iii. Indexing each file based on the primary keys.
9. Loading of data onto firebase, included the following tasks:
    i. Ensure that there is no pre-existing data. Deletion of data if it is already present on the firebase and loading data fresh.
    ii. Loading of clean data on firebase real-time database.

Screenshots of the Completed components:

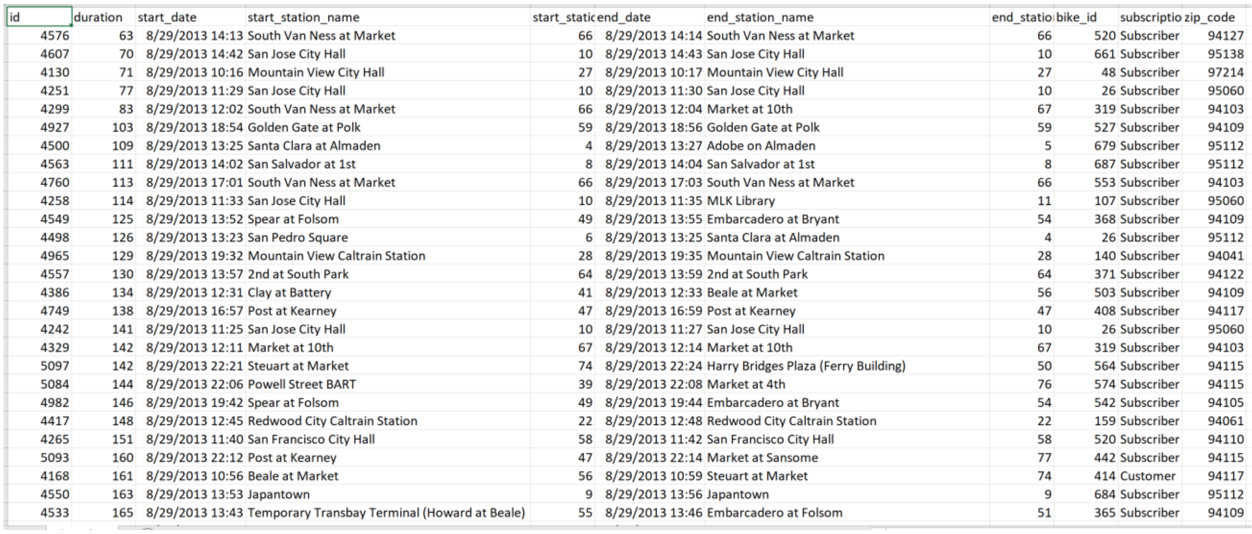
Sample of cleaned Kaggle Datasets:

station_clean:



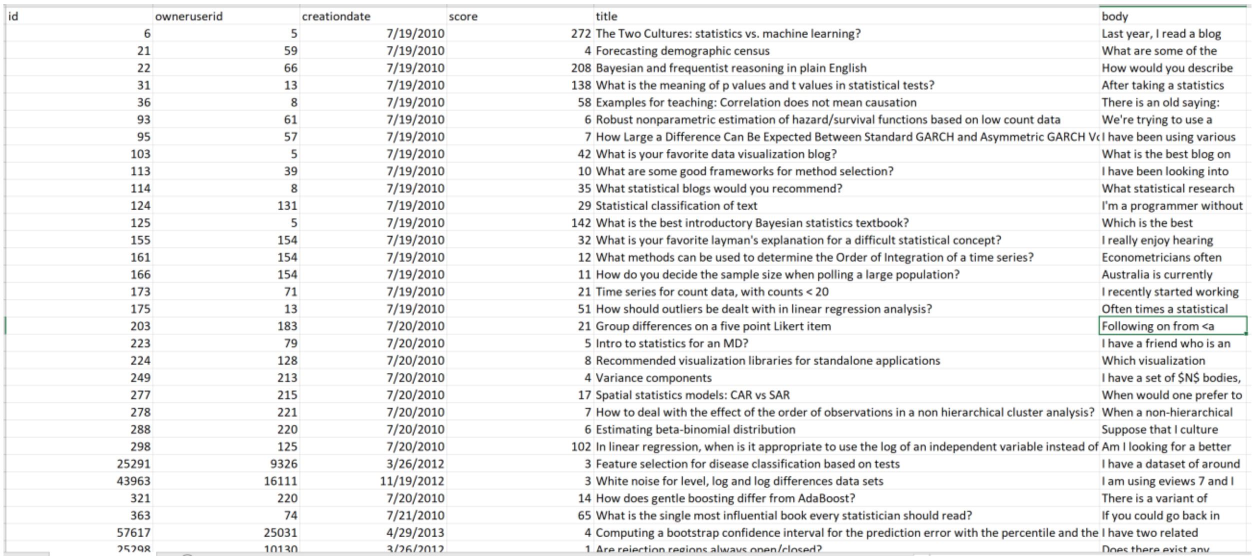
id	name	lat	long	dock_count	city	installation_date
2	San Jose Diridon Caltrain Station	37.329732	-121.901782	27	San Jose	8/6/2013
3	San Jose Civic Center	37.330698	-121.888979	15	San Jose	8/5/2013
4	Santa Clara at Almaden	37.333988	-121.894902	11	San Jose	8/6/2013
5	Adobe on Almaden	37.331415	-121.8932	19	San Jose	8/5/2013
6	San Pedro Square	37.336721	-121.894074	15	San Jose	8/7/2013
7	Paseo de San Antonio	37.333798	-121.886943	15	San Jose	8/7/2013
8	San Salvador at 1st	37.330165	-121.885831	15	San Jose	8/5/2013
9	Japantown	37.348742	-121.894715	15	San Jose	8/5/2013
10	San Jose City Hall	37.337391	-121.886995	15	San Jose	8/6/2013
11	MLK Library	37.335885	-121.88566	19	San Jose	8/6/2013
12	SJSU 4th at San Carlos	37.332808	-121.883891	19	San Jose	8/7/2013
13	St James Park	37.339301	-121.889937	15	San Jose	8/6/2013
14	Arena Green / SAP Center	37.332692	-121.900084	19	San Jose	8/5/2013
16	SJSU - San Salvador at 9th	37.333955	-121.877349	15	San Jose	8/7/2013
21	Franklin at Maple	37.481758	-122.226904	15	Redwood City	8/12/2013
22	Redwood City Caltrain Station	37.486078	-122.232089	25	Redwood City	8/15/2013
23	San Mateo County Center	37.487616	-122.229951	15	Redwood City	8/15/2013
24	Redwood City Public Library	37.484219	-122.227424	15	Redwood City	8/12/2013
25	Stanford in Redwood City	37.48537	-122.203288	15	Redwood City	8/12/2013
26	Redwood City Medical Center	37.487682	-122.223492	15	Redwood City	8/12/2013
27	Mountain View City Hall	37.389218	-122.081896	15	Mountain View	8/16/2013
28	Mountain View Caltrain Station	37.394358	-122.076713	23	Mountain View	8/15/2013
29	San Antonio Caltrain Station	37.40694	-122.106758	23	Mountain View	8/15/2013
30	Evelyn Park and Ride	37.390277	-122.066553	15	Mountain View	8/16/2013
31	San Antonio Shopping Center	37.400443	-122.108338	15	Mountain View	12/31/2013
32	Castro Street and El Camino Real	37.385956	-122.083678	11	Mountain View	12/31/2013
33	Rengstorff Avenue / California Street	37.400241	-122.099076	15	Mountain View	8/16/2013

trips_clean:

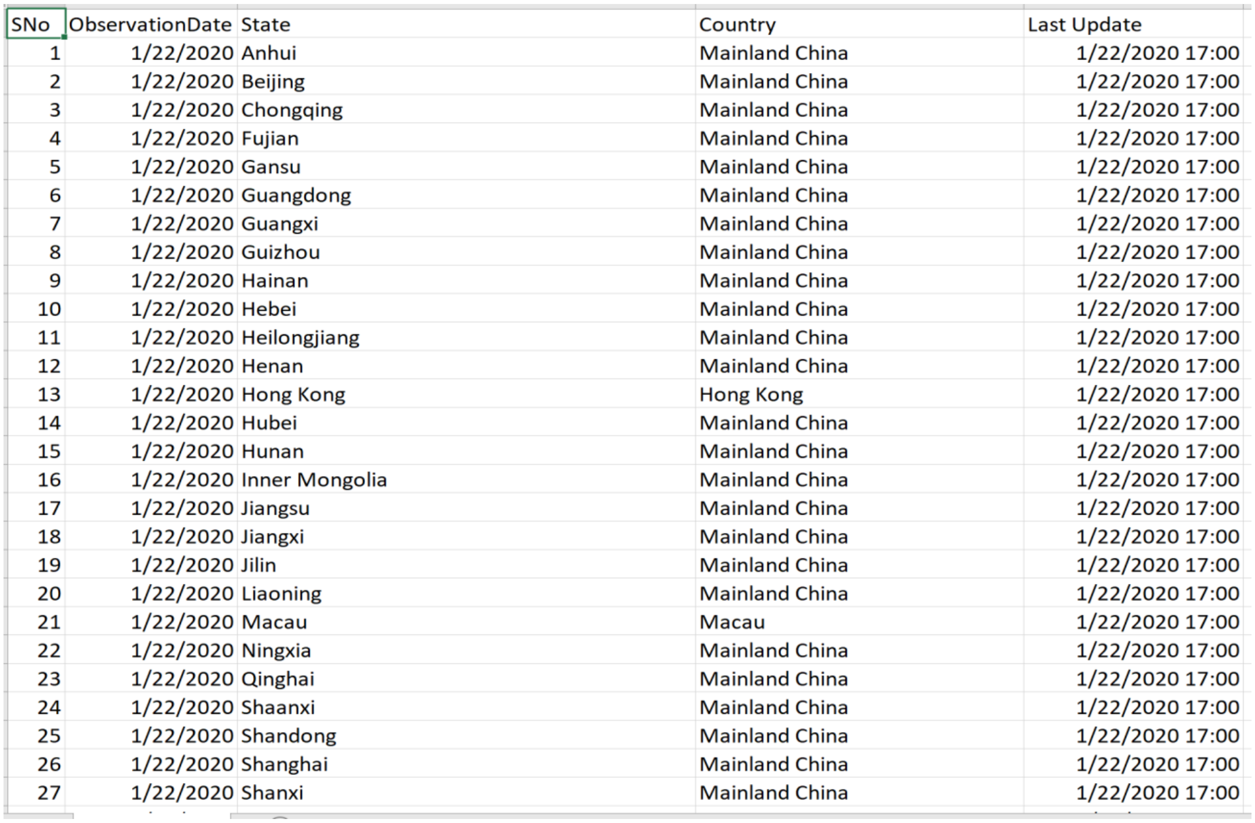


id	duration	start_date	start_station_name	start_static	end_date	end_station_name	end_statio	bike_id	subscriptio	zip_code
4576	63	8/29/2013 14:13	South Van Ness at Market	66	8/29/2013 14:14	South Van Ness at Market	66	520	Subscriber	94127
4607	70	8/29/2013 14:42	San Jose City Hall	10	8/29/2013 14:43	San Jose City Hall	10	661	Subscriber	95138
4130	71	8/29/2013 10:16	Mountain View City Hall	27	8/29/2013 10:17	Mountain View City Hall	27	48	Subscriber	97214
4251	77	8/29/2013 11:29	San Jose City Hall	10	8/29/2013 11:30	San Jose City Hall	10	26	Subscriber	95060
4299	83	8/29/2013 12:02	South Van Ness at Market	66	8/29/2013 12:04	Market at 10th	67	319	Subscriber	94103
4927	103	8/29/2013 18:54	Golden Gate at Polk	59	8/29/2013 18:56	Golden Gate at Polk	59	527	Subscriber	94109
4500	109	8/29/2013 13:25	Santa Clara at Almaden	4	8/29/2013 13:27	Adobe on Almaden	5	679	Subscriber	95112
4563	111	8/29/2013 14:02	San Salvador at 1st	8	8/29/2013 14:04	San Salvador at 1st	8	687	Subscriber	95112
4760	113	8/29/2013 17:01	South Van Ness at Market	66	8/29/2013 17:03	South Van Ness at Market	66	553	Subscriber	94103
4258	114	8/29/2013 11:33	San Jose City Hall	10	8/29/2013 11:35	MLK Library	11	107	Subscriber	95060
4549	125	8/29/2013 13:52	Spear at Folsom	49	8/29/2013 13:55	Embarcadero at Bryant	54	368	Subscriber	94109
4498	126	8/29/2013 13:23	San Pedro Square	6	8/29/2013 13:25	Santa Clara at Almaden	4	26	Subscriber	95112
4965	129	8/29/2013 19:32	Mountain View Caltrain Station	28	8/29/2013 19:35	Mountain View Caltrain Station	28	140	Subscriber	94041
4557	130	8/29/2013 13:57	2nd at South Park	64	8/29/2013 13:59	2nd at South Park	64	371	Subscriber	94122
4386	134	8/29/2013 12:31	Clay at Battery	41	8/29/2013 12:33	Beale at Market	56	503	Subscriber	94109
4749	138	8/29/2013 16:57	Post at Kearney	47	8/29/2013 16:59	Post at Kearney	47	408	Subscriber	94117
4242	141	8/29/2013 11:25	San Jose City Hall	10	8/29/2013 11:27	San Jose City Hall	10	26	Subscriber	95060
4329	142	8/29/2013 12:11	Market at 10th	67	8/29/2013 12:14	Market at 10th	67	319	Subscriber	94103
5097	142	8/29/2013 22:21	Steuart at Market	74	8/29/2013 22:24	Harry Bridges Plaza (Ferry Building)	50	564	Subscriber	94115
5084	144	8/29/2013 22:06	Powell Street BART	39	8/29/2013 22:08	Market at 4th	76	574	Subscriber	94115
4982	146	8/29/2013 19:42	Spear at Folsom	49	8/29/2013 19:44	Embarcadero at Bryant	54	542	Subscriber	94105
4417	148	8/29/2013 12:45	Redwood City Caltrain Station	22	8/29/2013 12:48	Redwood City Caltrain Station	22	159	Subscriber	94061
4265	151	8/29/2013 11:40	San Francisco City Hall	58	8/29/2013 11:42	San Francisco City Hall	58	520	Subscriber	94110
5093	160	8/29/2013 22:12	Post at Kearney	47	8/29/2013 22:14	Market at Sansome	77	442	Subscriber	94115
4168	161	8/29/2013 10:56	Beale at Market	56	8/29/2013 10:59	Steuart at Market	74	414	Customer	94117
4550	163	8/29/2013 13:53	Japantown	9	8/29/2013 13:56	Japantown	9	684	Subscriber	95112
4533	165	8/29/2013 13:43	Temporary Transbay Terminal (Howard at Beale)	55	8/29/2013 13:46	Embarcadero at Folsom	51	365	Subscriber	94109

questions_clean:

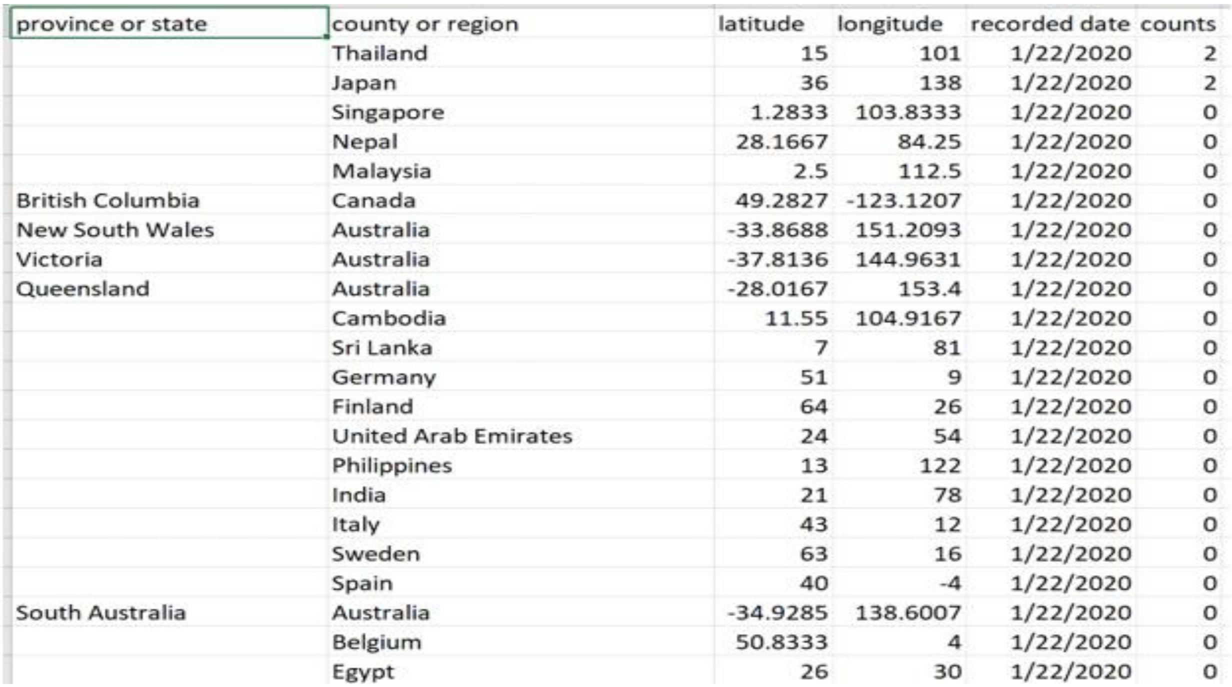


id	owneruserid	creationdate	score	title	body
6	5	7/19/2010	272	The Two Cultures: statistics vs. machine learning?	Last year, I read a blog
21	59	7/19/2010	4	Forecasting demographic census	What are some of the
22	66	7/19/2010	208	Bayesian and frequentist reasoning in plain English	How would you describe
31	13	7/19/2010	138	What is the meaning of p values and t values in statistical tests?	After taking a statistics
36	8	7/19/2010	58	Examples for teaching: Correlation does not mean causation	There is an old saying:
93	61	7/19/2010	6	Robust nonparametric estimation of hazard/survival functions based on low count data	We're trying to use a
95	57	7/19/2010	7	How Large a Difference Can Be Expected Between Standard GARCH and Asymmetric GARCH V	I have been using various
103	5	7/19/2010	42	What is your favorite data visualization blog?	What is the best blog on
113	39	7/19/2010	10	What are some good frameworks for method selection?	I have been looking into
114	8	7/19/2010	35	What statistical blogs would you recommend?	What statistical research
124	131	7/19/2010	29	Statistical classification of text	I'm a programmer without
125	5	7/19/2010	142	What is the best introductory Bayesian statistics textbook?	Which is the best
155	154	7/19/2010	32	What is your favorite layman's explanation for a difficult statistical concept?	I really enjoy hearing
161	154	7/19/2010	12	What methods can be used to determine the Order of Integration of a time series?	Econometricians often
166	154	7/19/2010	11	How do you decide the sample size when polling a large population?	Australia is currently
173	71	7/19/2010	21	Time series for count data, with counts < 20	I recently started working
175	13	7/19/2010	51	How should outliers be dealt with in linear regression analysis?	Often times a statistical
203	183	7/20/2010	21	Group differences on a five point Likert item	Following on from <a
223	79	7/20/2010	5	Intro to statistics for an MD?	I have a friend who is an
224	128	7/20/2010	8	Recommended visualization libraries for standalone applications	Which visualization
249	213	7/20/2010	4	Variance components	I have a set of N bodies,
277	215	7/20/2010	17	Spatial statistics models: CAR vs SAR	When would one prefer to
278	221	7/20/2010	7	How to deal with the effect of the order of observations in a non hierarchical cluster analysis?	When a non-hierarchical
288	220	7/20/2010	6	Estimating beta-binomial distribution	Suppose that I culture
298	125	7/20/2010	102	In linear regression, when is it appropriate to use the log of an independent variable instead of	Am I looking for a better
25291	9326	3/26/2012	3	Feature selection for disease classification based on tests	I have a dataset of around
43963	16111	11/19/2012	3	White noise for level, log and log differences data sets	I am using eviews 7 and I
321	220	7/20/2010	14	How does gentle boosting differ from AdaBoost?	There is a variant of
363	74	7/21/2010	65	What is the single most influential book every statistician should read?	If you could go back in
57617	25031	4/29/2013	4	Computing a bootstrap confidence interval for the prediction error with the percentile and the	I have two related
25298	10130	3/26/2012	1	Are rejection regions always open/closed?	Does there exist any



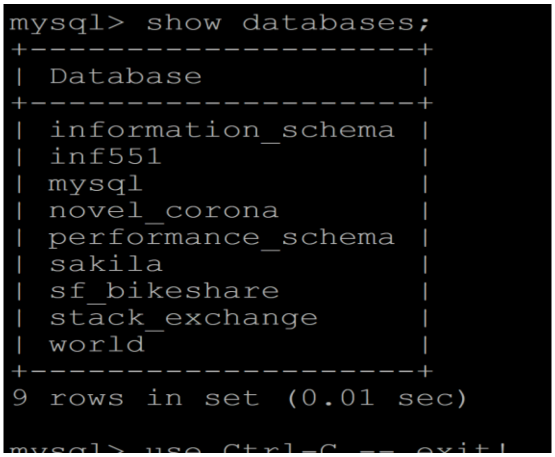
SNo	ObservationDate	State	Country	Last Update
1	1/22/2020	Anhui	Mainland China	1/22/2020 17:00
2	1/22/2020	Beijing	Mainland China	1/22/2020 17:00
3	1/22/2020	Chongqing	Mainland China	1/22/2020 17:00
4	1/22/2020	Fujian	Mainland China	1/22/2020 17:00
5	1/22/2020	Gansu	Mainland China	1/22/2020 17:00
6	1/22/2020	Guangdong	Mainland China	1/22/2020 17:00
7	1/22/2020	Guangxi	Mainland China	1/22/2020 17:00
8	1/22/2020	Guizhou	Mainland China	1/22/2020 17:00
9	1/22/2020	Hainan	Mainland China	1/22/2020 17:00
10	1/22/2020	Hebei	Mainland China	1/22/2020 17:00
11	1/22/2020	Heilongjiang	Mainland China	1/22/2020 17:00
12	1/22/2020	Henan	Mainland China	1/22/2020 17:00
13	1/22/2020	Hong Kong	Hong Kong	1/22/2020 17:00
14	1/22/2020	Hubei	Mainland China	1/22/2020 17:00
15	1/22/2020	Hunan	Mainland China	1/22/2020 17:00
16	1/22/2020	Inner Mongolia	Mainland China	1/22/2020 17:00
17	1/22/2020	Jiangsu	Mainland China	1/22/2020 17:00
18	1/22/2020	Jiangxi	Mainland China	1/22/2020 17:00
19	1/22/2020	Jilin	Mainland China	1/22/2020 17:00
20	1/22/2020	Liaoning	Mainland China	1/22/2020 17:00
21	1/22/2020	Macau	Macau	1/22/2020 17:00
22	1/22/2020	Ningxia	Mainland China	1/22/2020 17:00
23	1/22/2020	Qinghai	Mainland China	1/22/2020 17:00
24	1/22/2020	Shaanxi	Mainland China	1/22/2020 17:00
25	1/22/2020	Shandong	Mainland China	1/22/2020 17:00
26	1/22/2020	Shanghai	Mainland China	1/22/2020 17:00
27	1/22/2020	Shanxi	Mainland China	1/22/2020 17:00

confirmed_cases:



province or state	county or region	latitude	longitude	recorded date	counts
	Thailand	15	101	1/22/2020	2
	Japan	36	138	1/22/2020	2
	Singapore	1.2833	103.8333	1/22/2020	0
	Nepal	28.1667	84.25	1/22/2020	0
	Malaysia	2.5	112.5	1/22/2020	0
British Columbia	Canada	49.2827	-123.1207	1/22/2020	0
New South Wales	Australia	-33.8688	151.2093	1/22/2020	0
Victoria	Australia	-37.8136	144.9631	1/22/2020	0
Queensland	Australia	-28.0167	153.4	1/22/2020	0
	Cambodia	11.55	104.9167	1/22/2020	0
	Sri Lanka	7	81	1/22/2020	0
	Germany	51	9	1/22/2020	0
	Finland	64	26	1/22/2020	0
	United Arab Emirates	24	54	1/22/2020	0
	Philippines	13	122	1/22/2020	0
	India	21	78	1/22/2020	0
	Italy	43	12	1/22/2020	0
	Sweden	63	16	1/22/2020	0
	Spain	40	-4	1/22/2020	0
South Australia	Australia	-34.9285	138.6007	1/22/2020	0
	Belgium	50.8333	4	1/22/2020	0
	Egypt	26	30	1/22/2020	0

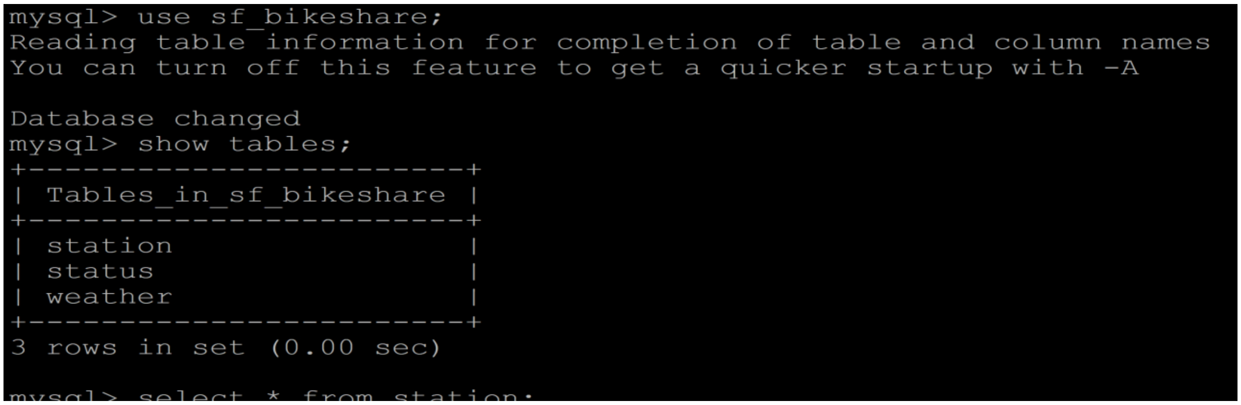
MySQL DB Data:



```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| inf551             |
| mysql              |
| novel_corona       |
| performance_schema |
| sakila             |
| sf_bikeshare       |
| stack_exchange     |
| world              |
+--------------------+
9 rows in set (0.01 sec)

mysql> use Ctrl-C -- exit!
```

SF- Bike share DB:

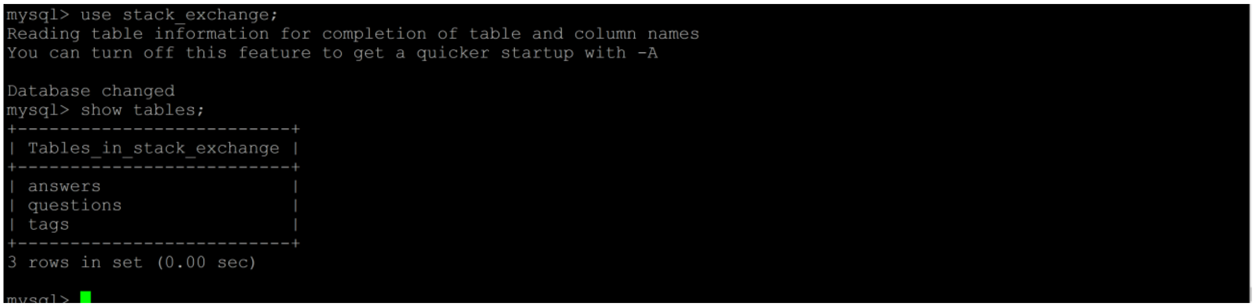


```
mysql> use sf_bikeshare;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+------------------------+
| Tables_in_sf_bikeshare |
+------------------------+
| station                |
| status                 |
| weather                |
+------------------------+
3 rows in set (0.00 sec)

mysql> select * from station;
```

Stack Exchange Database

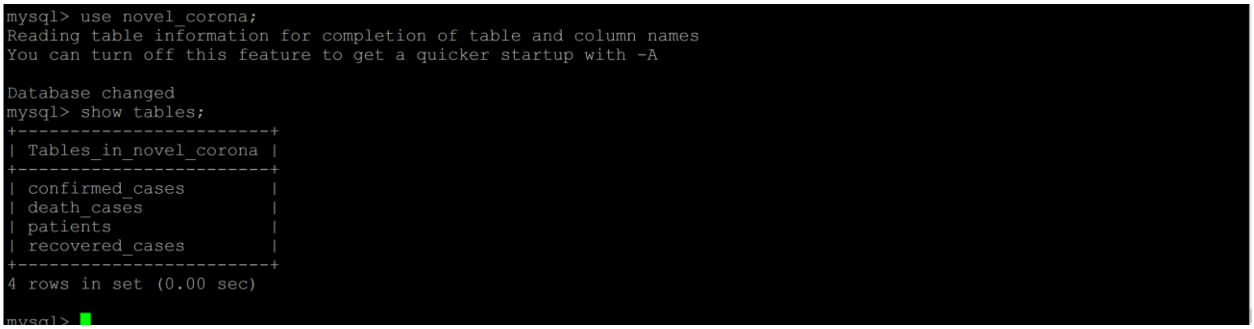


```
mysql> use stack_exchange;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+--------------------------+
| Tables_in_stack_exchange |
+--------------------------+
| answers                  |
| questions                |
| tags                     |
+--------------------------+
3 rows in set (0.00 sec)

mysql>
```

Novel Corona Database:

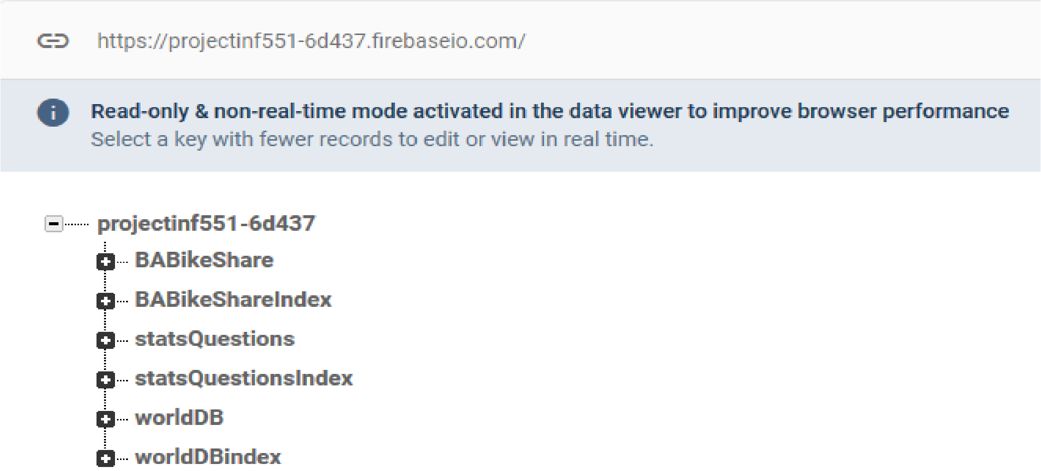


```
mysql> use novel_corona;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+------------------------+
| Tables_in_novel_corona |
+------------------------+
| confirmed_cases        |
| death_cases            |
| patients               |
| recovered_cases        |
+------------------------+
4 rows in set (0.00 sec)

mysql>
```

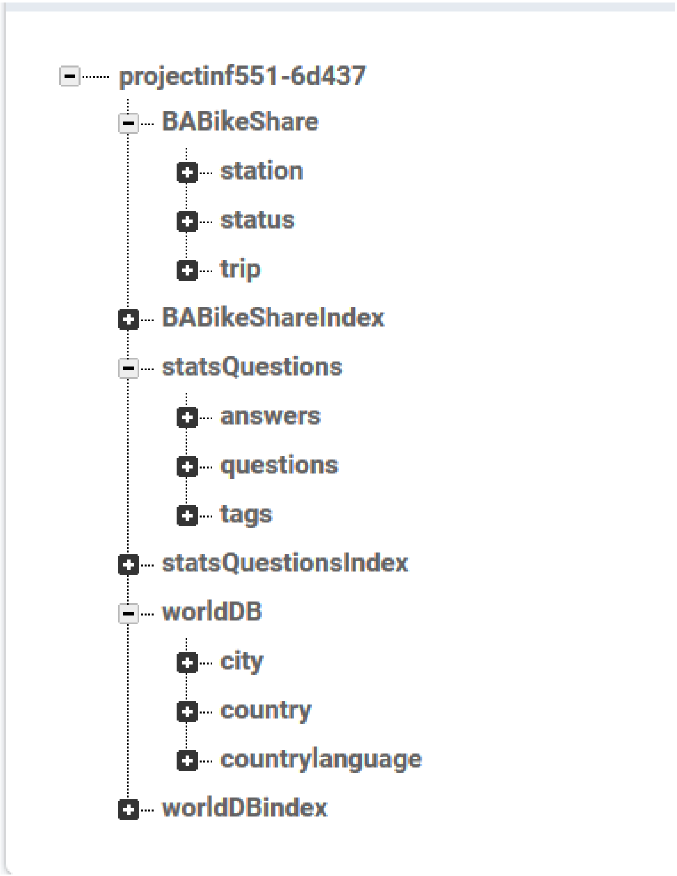
Data loaded in Firebase:



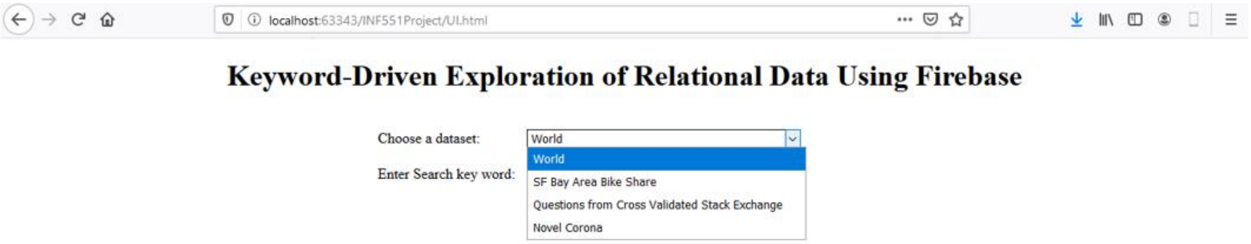
https://projectinf551-6d437.firebaseio.com/

Read-only & non-real-time mode activated in the data viewer to improve browser performance
Select a key with fewer records to edit or view in real time.

- projectinf551-6d437
  - BABikeShare
  - BABikeShareIndex
  - statsQuestions
  - statsQuestionsIndex
  - worldDB
  - worldDBindex



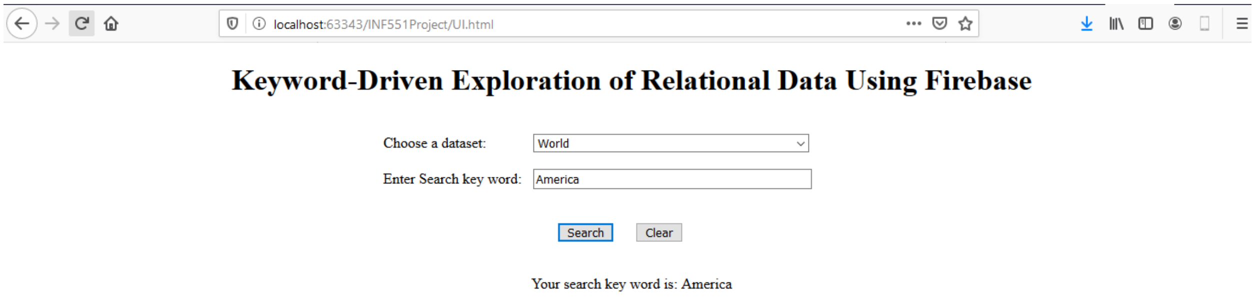
Basic UI:


localhost:63343/INF551Project/UI.html
World
SF Bay Area Bike Share
Questions from Cross Validated Stack Exchange
Novel Corona

Keyword-Driven Exploration of Relational Data Using Firebase

Choose a dataset: World

Enter Search key word:



Keyword-Driven Exploration of Relational Data Using Firebase

Choose a dataset: World

Enter Search key word: America

Your search key word is: America

Challenges encountered:

1. Handling large datasets was an issue due to file size being large to open on local file systems
2. Cleansing of large data sets was time consuming
3. Deciding on to a single format to store and access data
4. Generalizing the code to work on all datasets