Project Lab Tutorial
Group B

Riki Chang, Carolina Rivera, David Sayad, Chris Wallace
California State University Los Angeles
E-mail:cwallac9@calstatela.edu, criver47@calstatela.edu, dsayad2@calstatela.edu, rchang12@calstatela.edu

Lending Club Loan Data Analysis

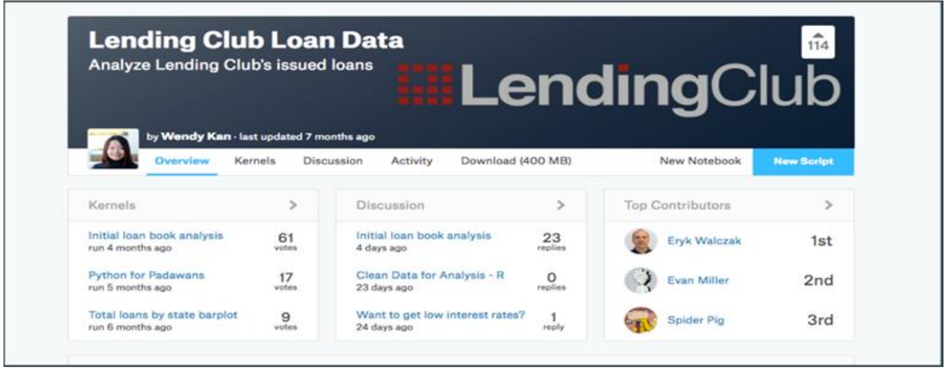
Objectives

In this hands-on lab, you will learn how to:

- Create IBM Bluemix Account
- Create directories in cluster and load data
- Learn how to use Ambari
- Introduced to Hive Shell
- Hive commands to perform the analysis.
- Visualization both Temporal and Geospatial

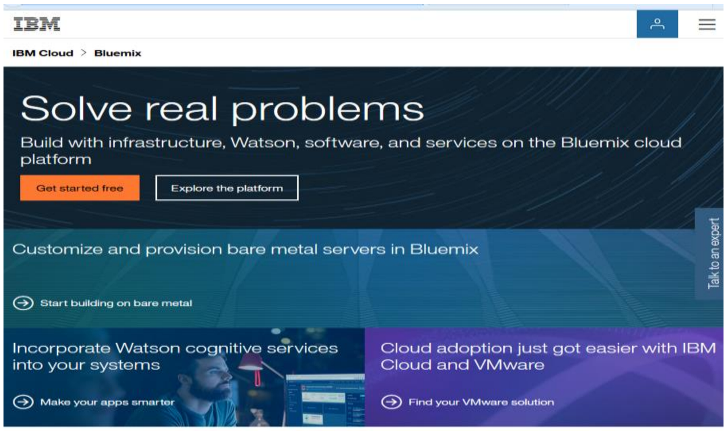
Exercise 1: Get data manually from keggle

1. Go to https://www.kaggle.com/wendykan/lending-club-loan-data to download the data and save it to your local machine. (Remember the location you save it too you will need it later)

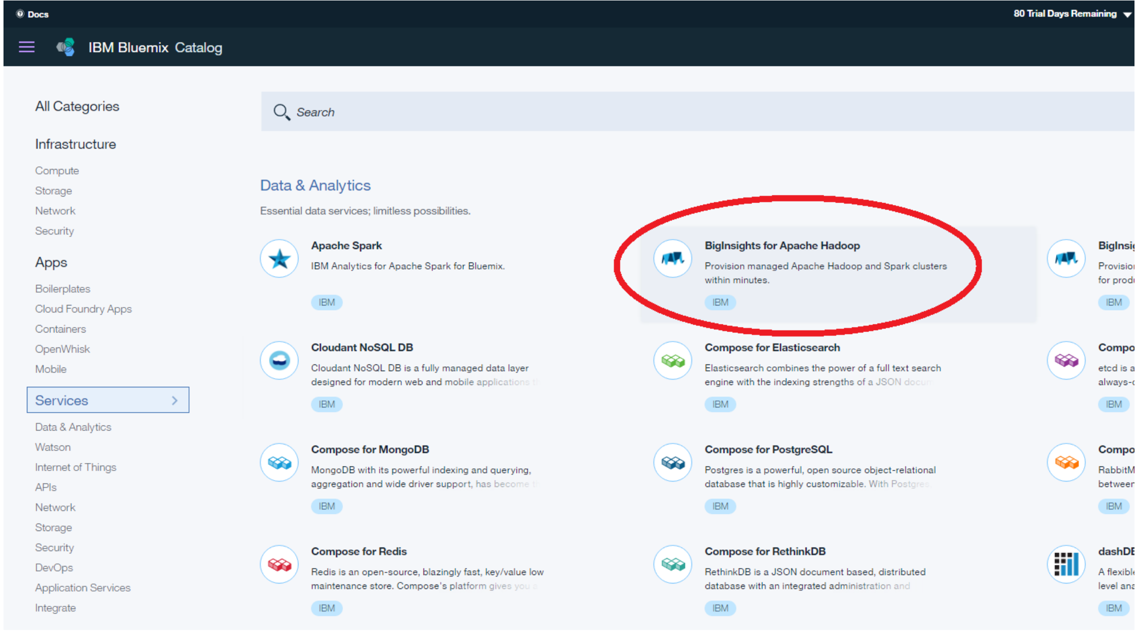


Exercise 2: Create Bluemix Account and a Cluster

1. Next open up a new tab in your browser and go to https://www.ibm.com/cloud-computing/bluemix/ to create your account.


IBM
IBM Cloud > Bluemix
Solve real problems
Build with infrastructure, Watson, software, and services on the Bluemix cloud platform
Get started free
Explore the platform
Customize and provision bare metal servers in Bluemix
Start building on bare metal
Talk to an expert
Incorporate Watson cognitive services into your systems
Make your apps smarter
Cloud adoption just got easier with IBM Cloud and VMware
Find your VMware solution

2. Select create service and select BigInSights for Apache Hadoop



Data & Analytics

Essential data services; limitless possibilities.

Apache Spark IBM Analytics for Apache Spark for Bluemix. IBM

BigInsights for Apache Hadoop Provision managed Apache Hadoop and Spark clusters within minutes. IBM

BigInsi Provisio for prod IBM

Cloudant NoSQL DB Cloudant NoSQL DB is a fully managed data layer designed for modern web and mobile applications t IBM

Compose for Elasticsearch Elasticsearch combines the power of a full text search engine with the indexing strengths of a JSON docum IBM

Compo etcd is a always-c IBM

Compose for MongoDB MongoDB with its powerful indexing and querying, aggregation and wide driver support, has become t IBM

Compose for PostgreSQL Postgres is a powerful, open source object-relational database that is highly customizable. With Postgres IBM

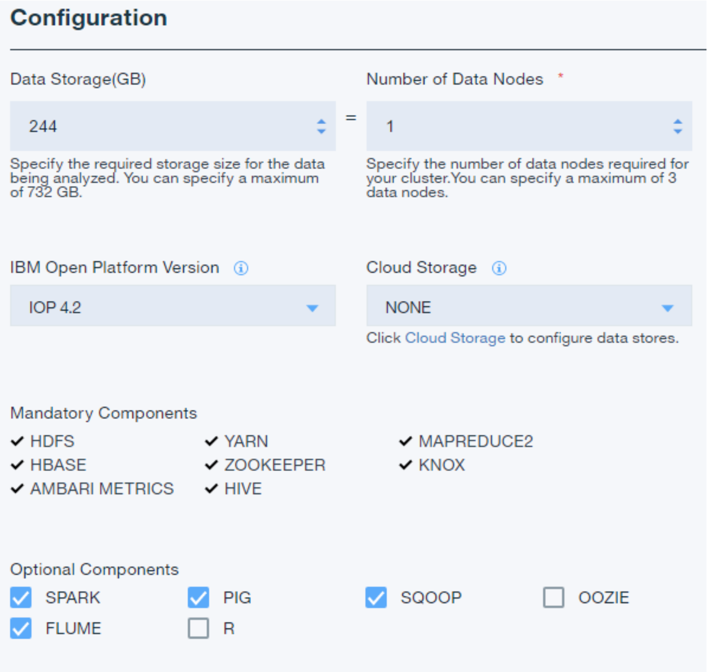
Compo RabbitM between IBM

Compose for Redis Redis is an open-source, blazingly fast, key/value low maintenance store. Compose's platform gives you a IBM

Compose for RethinkDB RethinkDB is a JSON document based, distributed database with an integrated administration and IBM

dashDE A flexible level ana IBM

3. Then navigate to **Manage Cluster** then **New Cluster** to create your cluster.



Configuration

Data Storage(GB): 244 = Number of Data Nodes *: 1

Specify the required storage size for the data being analyzed. You can specify a maximum of 732 GB.

Specify the number of data nodes required for your cluster.You can specify a maximum of 3 data nodes.

IBM Open Platform Version: IOP 4.2

Cloud Storage: NONE

Click Cloud Storage to configure data stores.

Mandatory Components

✓ HDFS ✓ YARN ✓ MAPREDUCE2
✓ HBASE ✓ ZOOKEEPER ✓ KNOX
✓ AMBARI METRICS ✓ HIVE

Optional Components

☑ SPARK ☑ PIG ☑ SQOOP ☐ OOZIE
☑ FLUME ☐ R

4. Make sure that **SPARK**, **PIG**, **SQOOP**, and **FLUME** are all checked off on the configuration and create your cluster

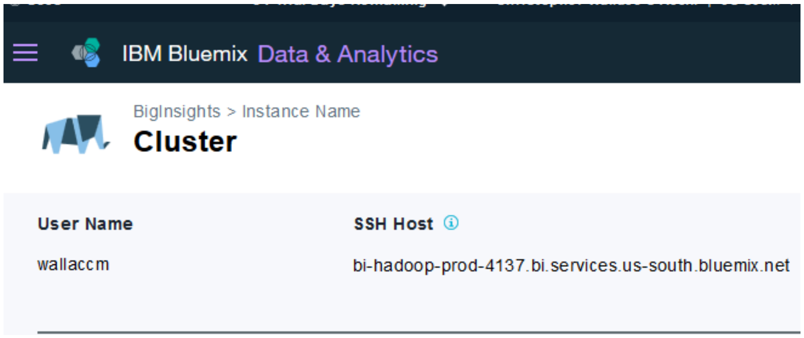
---

Exercise 3: Create directories in cluster and load data

---

1. Open up terminal/Putty on your the machine you are using and we will login to HDFS by typing in this command substituted with your info

$ssh username@bi-hadoop-prod-4137.bi.services.us-south.bluemix.net



IBM Bluemix Data & Analytics

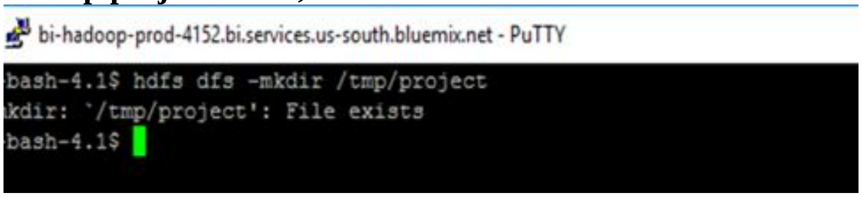
BigInsights > Instance Name

Cluster

User Name	SSH Host
wallaccm	bi-hadoop-prod-4137.bi.services.us-south.bluemix.net

2. After logging in enter the next command to create a directory inside HDFS that will be used to store the data that we downloaded from keggle in the previous step.

$ hdfs dfs -mkdir tmp/project;
$ hdfs dfs -mkdir tmp/project/tables;



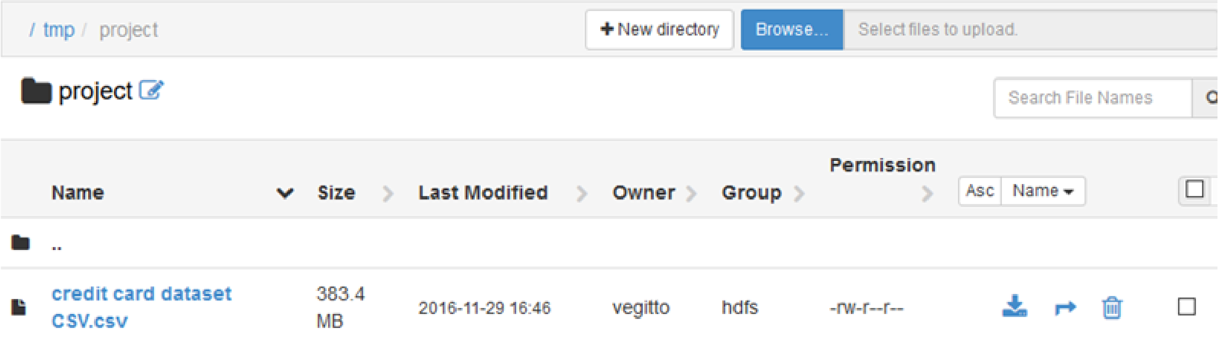
bi-hadoop-prod-4152.bi.services.us-south.bluemix.net - PuTTY

```
bash-4.1$ hdfs dfs -mkdir /tmp/project
kdir: `/tmp/project': File exists
bash-4.1$
```

3. Next we will manually upload the data through Ambari. Navigate to the same page where you grabbed your login link and click on the link entitled Ambari. Use the same username and password that you used to log into HDFS. After logging into Ambari do as follows:

FileBrowser > User > (username) > tmp > project > tables

4. Then select **Upload** on the top right and upload the data.



Name	Size	Last Modified	Owner	Group	Permission	
..						
credit card dataset CSV.csv	383.4 MB	2016-11-29 16:46	vegitto	hdfs	-rw-r--r--	□

Exercise 4: Hive Commands

1. Start up your hive shell by entering the following command.

$ hive

2. In the hive shell we will upload the data into a table stored in the tables directory. Next we will create a table that will extract the columns of the data that we are looking for.

hive>CREATE EXTERNAL TABLE IF NOT Exists project(json_responce STRING) STORED AS TEXTFILE LOCATION "/tmp/project/tables";

hive>CREATE TABLE IF NOT EXISTS project(id BIGINT, loan_amnt BIGINT, home_ownership STRING, issue_d STRING, loan_status STRING, purpose STRING, addr_state STRING);

INSERT OVERWRITE TABLE project SELECT id, loan_amnt, home_ownership, issue_d, loan_status, purpose, addr_state FROM credit_card_dataset_CSV.csv WHERE id = id

3. To be sure that both command work type in the following command to look at all the tables in your hive shell.

hive> show tables;

4. Once it is verified that the tables have been created you can execute a number of commands to query the table further.

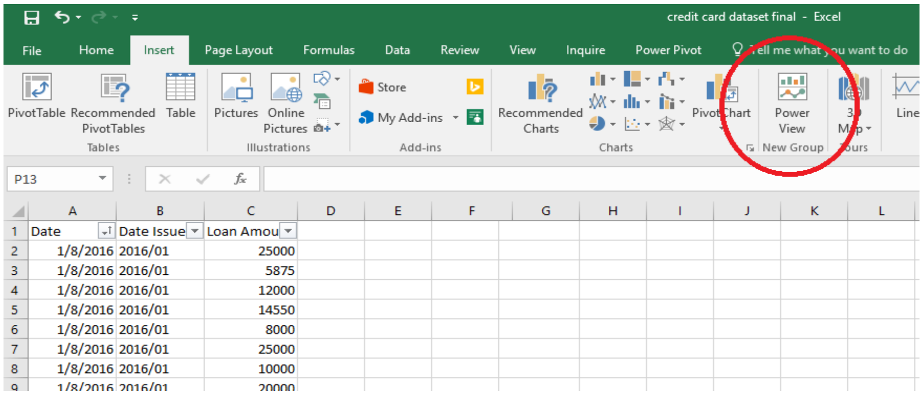
Example:
hive> Select * From project limit 50;
Which translates to select the first fifty rows from the project table.

Exercise 5: Temporal Analysis

1. To visualize the data, you must download the csv file from Ambari File Manager.

2. From there you open this file in excel and save it as an excel file(xlsx).

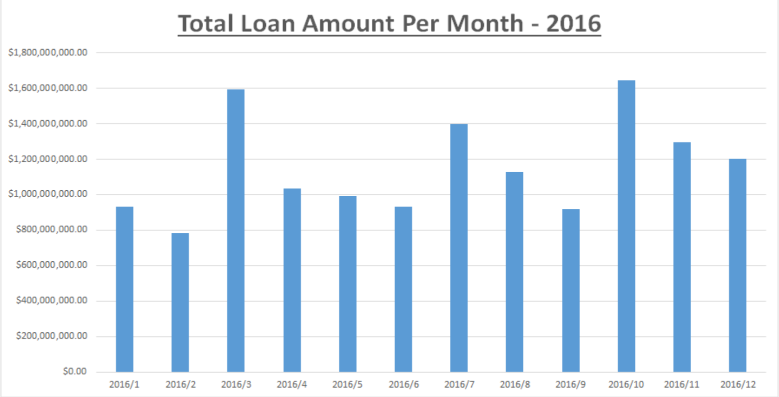
3. From there navigate to the Insert tab and select Power View to begin a Power View Report. (Make sure that you enable Power View through the options under all add-ins and add it to the Excel Ribbon)



	A	B	C	D	E	F	G	H	I	J	K	L
1	Date	Date Issue	Loan Amou									
2	1/8/2016	2016/01	25000									
3	1/8/2016	2016/01	5875									
4	1/8/2016	2016/01	12000									
5	1/8/2016	2016/01	14550									
6	1/8/2016	2016/01	8000									
7	1/8/2016	2016/01	25000									
8	1/8/2016	2016/01	10000									
9	1/8/2016	2016/01	20000									

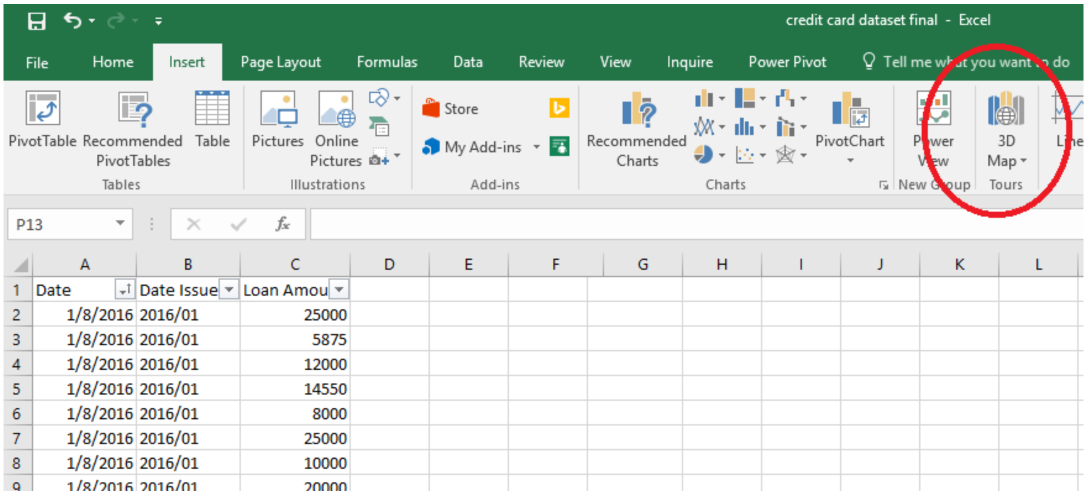
4. From here you select your Power View Fields that you would like to visualize (date, loan amount). Make sure to order and format the date in order to output the correct graph.

5. Next from the Design Tab select Other Charts (Bar) to display a bar graph of Total Loan Amounts given by the Leading Club shown below.



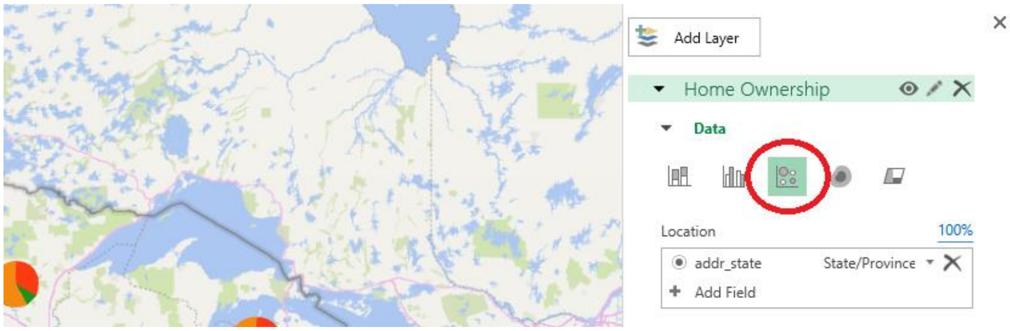
Exercise 6: Geospatial Analysis

1. First you must have the Data Analysis add-in enabled in excel. After making sure that it is enabled, navigate to the Insert Tab and select **3D Map.**

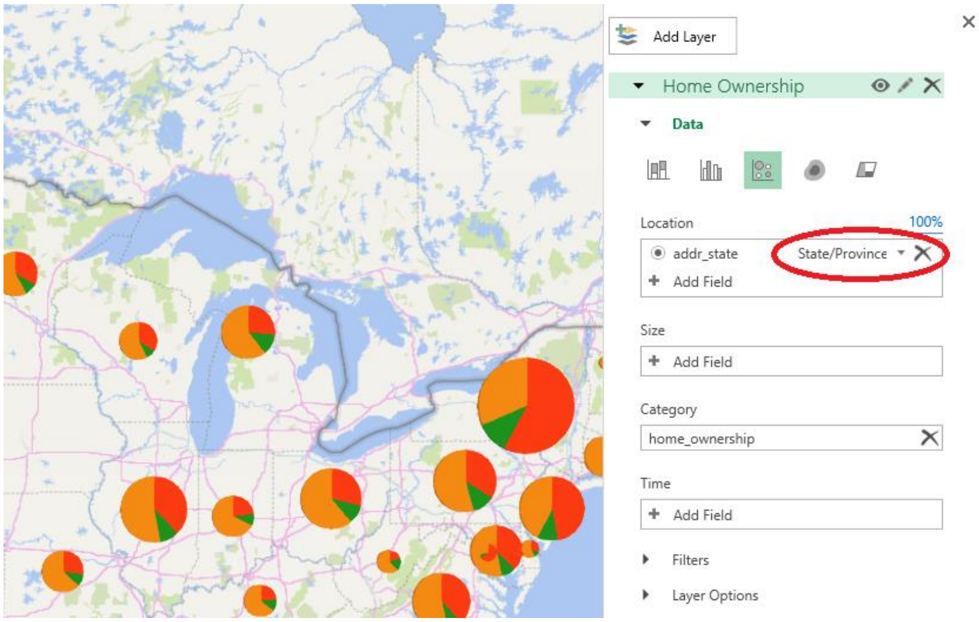


	A	B	C	D	E	F	G	H	I	J	K	L
1	Date	Date Issue	Loan Amou									
2	1/8/2016	2016/01	25000									
3	1/8/2016	2016/01	5875									
4	1/8/2016	2016/01	12000									
5	1/8/2016	2016/01	14550									
6	1/8/2016	2016/01	8000									
7	1/8/2016	2016/01	25000									
8	1/8/2016	2016/01	10000									
9	1/8/2016	2016/01	20000									

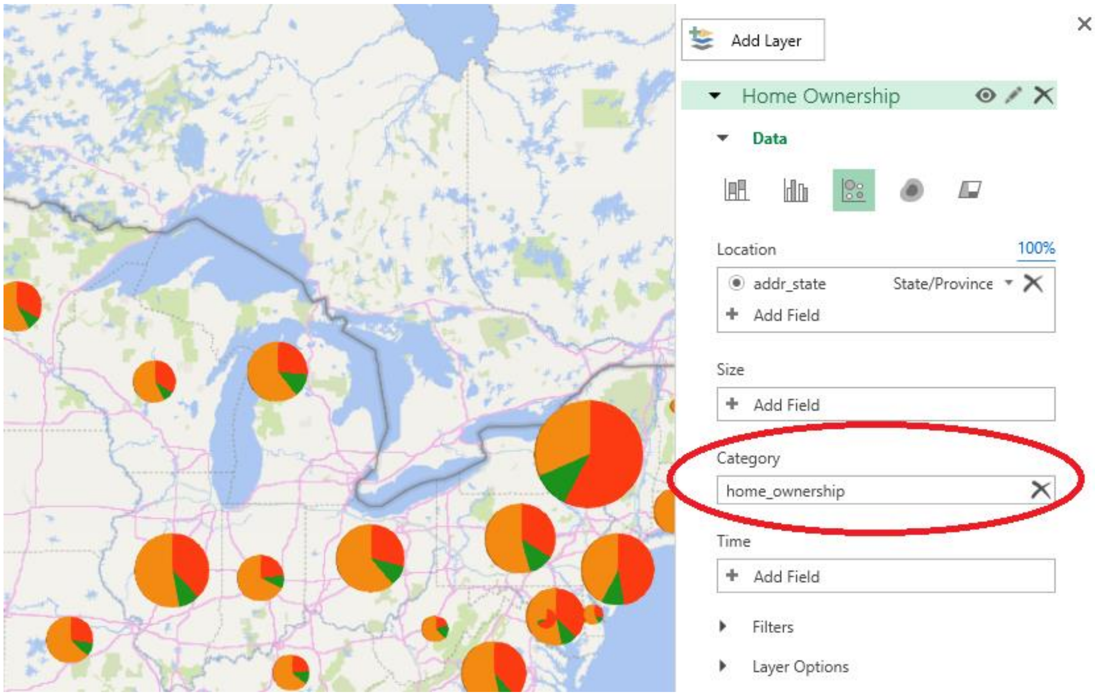
2. From here you must create different layers to show different attributes of data. Select **Add Layer** and select **Bubble** visualization.



3. From here the **Location** of the data will always be *addr_state* and you must select **State/Province.**



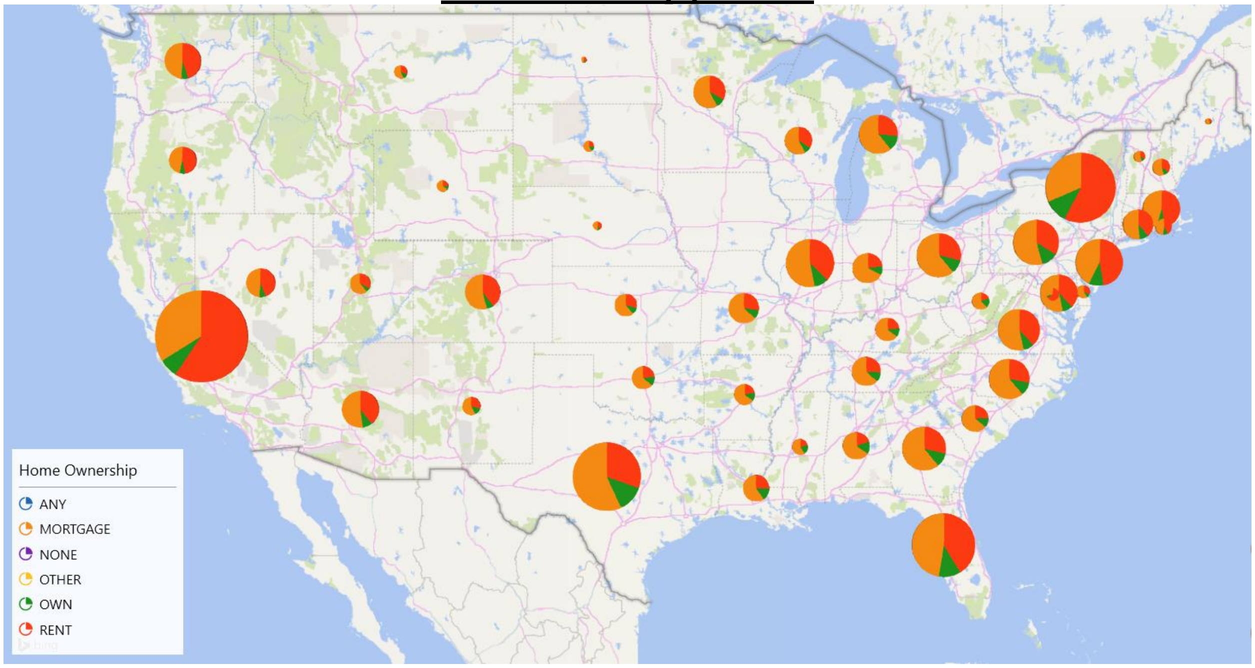
4. Then for **Category** select the attribute you would like to to visualize. We selected to have it in terms of Home Ownership, Loan Status, and Purpose. (Shown Below)



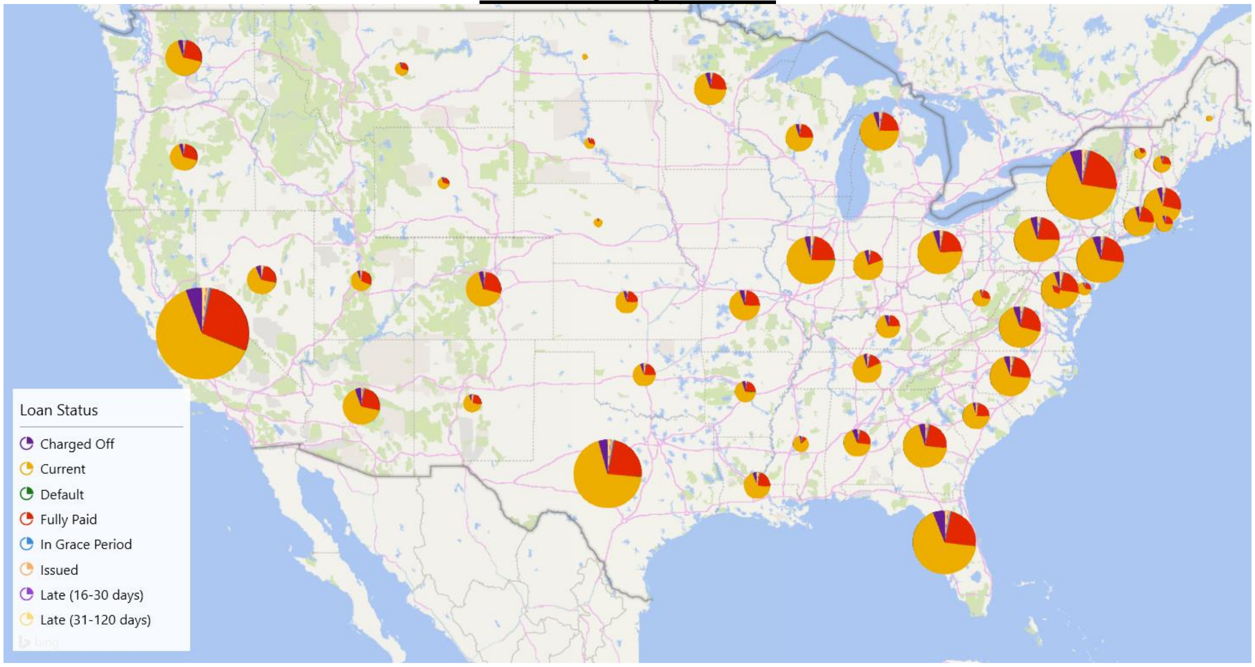
Location 100%
addr_state State/Province

Category
home_ownership

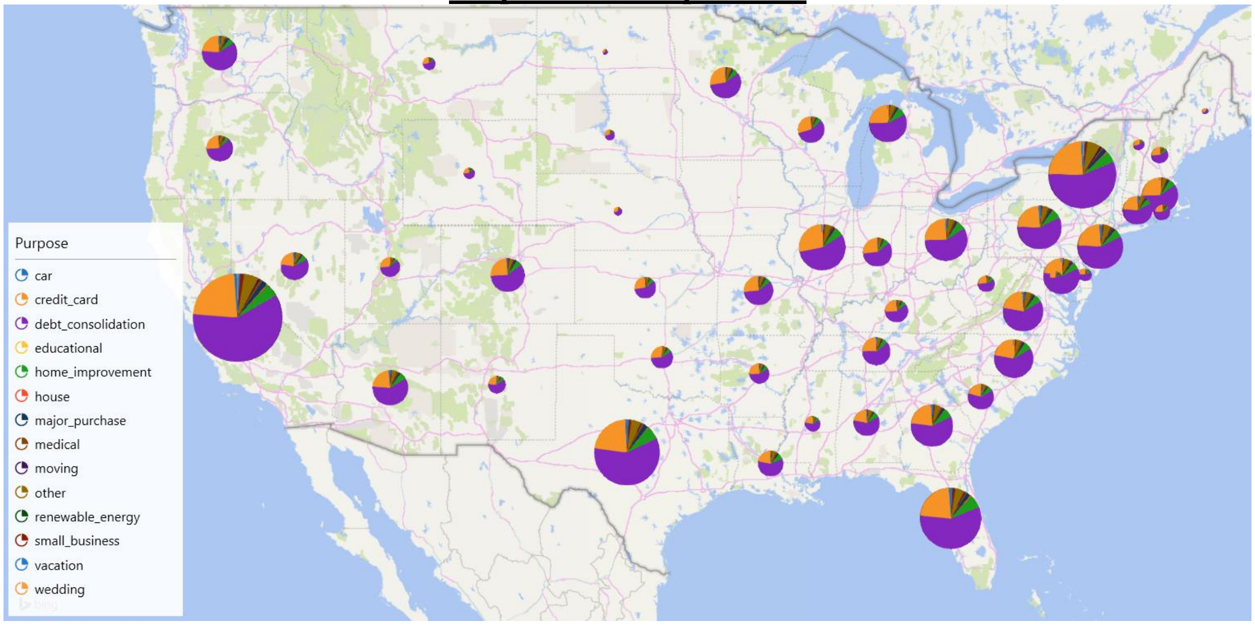
Home Ownership per State



Loan Status per State



Purpose of Loan per State



END OF LAB