DS-GA-1007 Final Project: NYC Taxi Analysis

Team Members:

Shixin Li - sl3368

Yili Yu - yy1734

Osvaldo Bulos - obr214

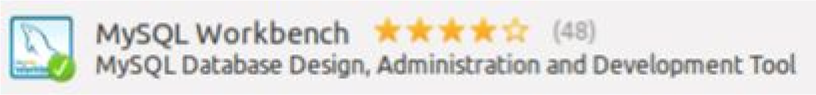
Introduction:

The goal of the project is to analyze the destinations of the Yellow Cabs given an origin or pick up location and a date. The project was developed under the Django framework, giving us great flexibility in the integration of other visualization tools such as HTML and Google Charts.

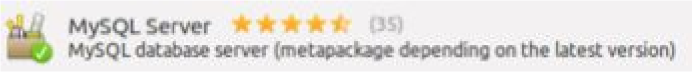
How to Run:

1. Open VirtualBox

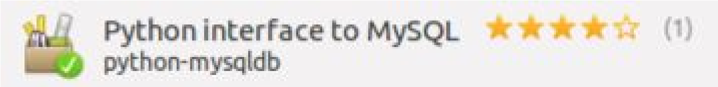
2. Download MySQL Workbench, MySQL Server and Python interface to MySQL from Ubuntu Software Center



MySQL Workbench ★★★★☆ (48)
MySQL Database Design, Administration and Development Tool

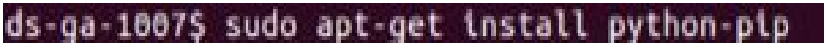


MySQL Server ★★★★★ (35)
MySQL database server (metapackage depending on the latest version)



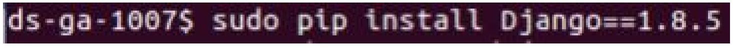
Python interface to MySQL ★★★★☆ (1)
python-mysqldb

3. Install pip:



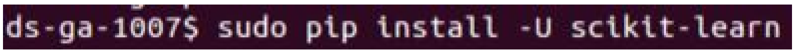
```
ds-ga-1007$ sudo apt-get install python-pip
```

4. Install Django:



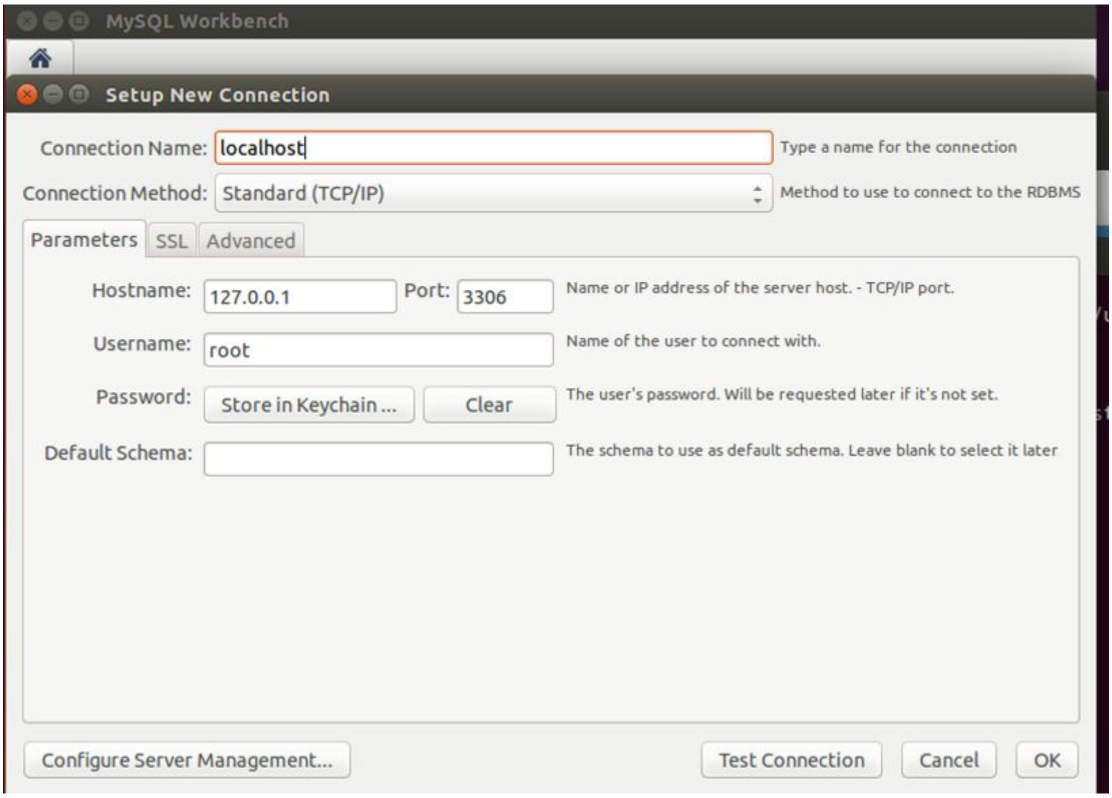
```
ds-ga-1007$ sudo pip install Django==1.8.5
```

5. Download Sklearn:



```
ds-ga-1007$ sudo pip install -U scikit-learn
```

6. Open MySQL Workbench. Type “localhost” in Connection Name.



Connection Name: localhost — Type a name for the connection

Connection Method: Standard (TCP/IP) — Method to use to connect to the RDBMS

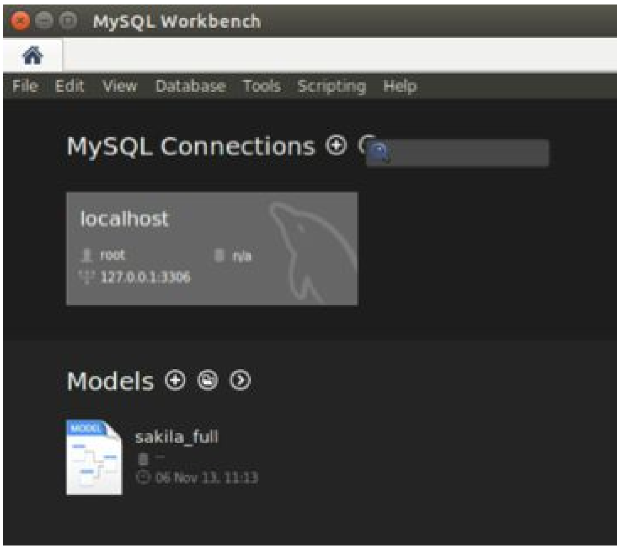
Hostname: 127.0.0.1 Port: 3306 — Name or IP address of the server host. - TCP/IP port.

Username: root — Name of the user to connect with.

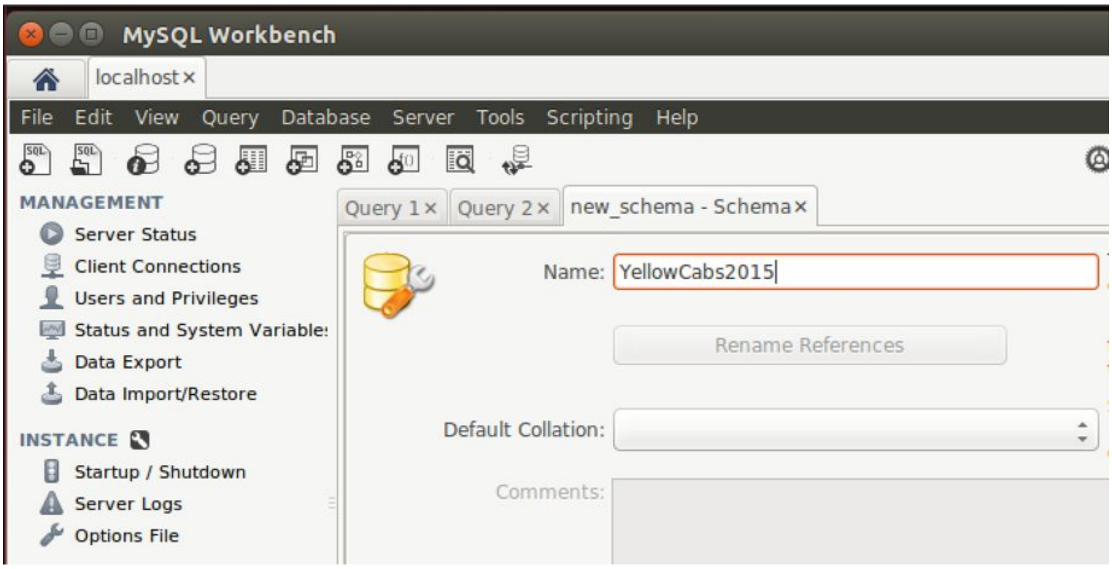
Password: — The user's password. Will be requested later if it's not set.

Default Schema: — The schema to use as default schema. Leave blank to select it later

7. Double Click localhost:



8. Click ‘Create a new schema’ and type ‘YellowCabs2015’ in the box

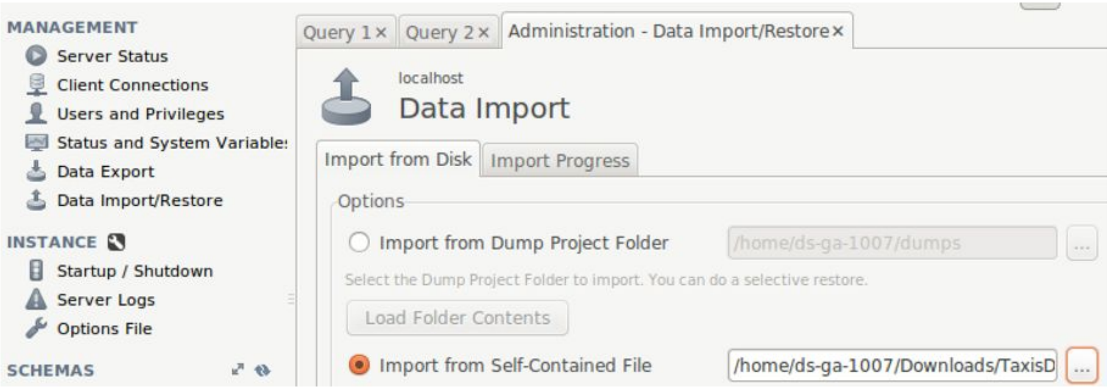


Name: YellowCabs2015

Default Collation:

9. Download TaxisDB.sql from the link below:
https://www.dropbox.com/s/4swreu0k0yosiiv/TaxisDB.sql?dl=0

10. Click ‘Data Import/Restore’. Choose “Import from Self-Contained File” and change the address to the folder where TaxisDB.sql from the previous step is saved. Then click “Start Import”

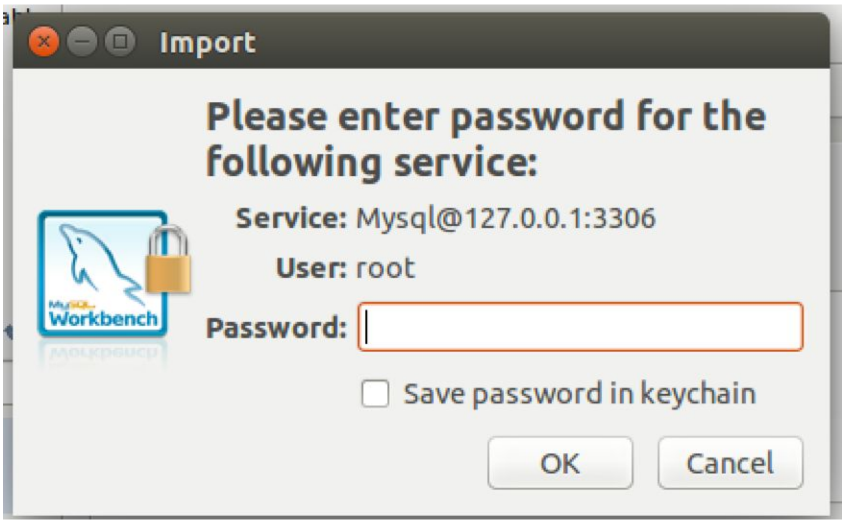


○ Import from Dump Project Folder /home/ds-ga-1007/dumps

Select the Dump Project Folder to import. You can do a selective restore.

● Import from Self-Contained File /home/ds-ga-1007/Downloads/TaxisD

11. A window will pop up asking for password after you click import. Keep the password box blank and click “OK”



Please enter password for the following service:

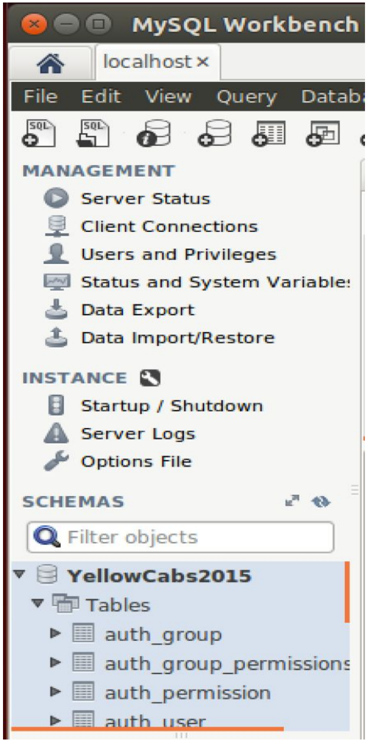
Service: Mysql@127.0.0.1:3306

User: root

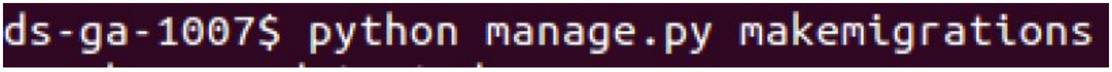
Password:

□ Save password in keychain

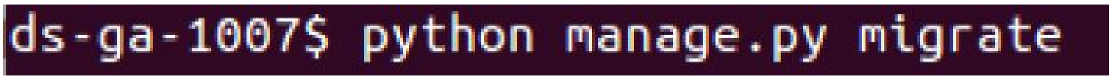
12. After Import Completed, on the left bar schemas section, select YellowCabs2015 --> Tables à taxis_taxipickups (scroll down to the bottom). Then right click on "taxis_taxipickups" and select "Select Rows – Limit 1000"



13. Open Terminal and direct the location to the folder where the code download from GitHub is saved. Type 'python manage.py runserver' and you may run into errors. Then you need to type the following two lines:

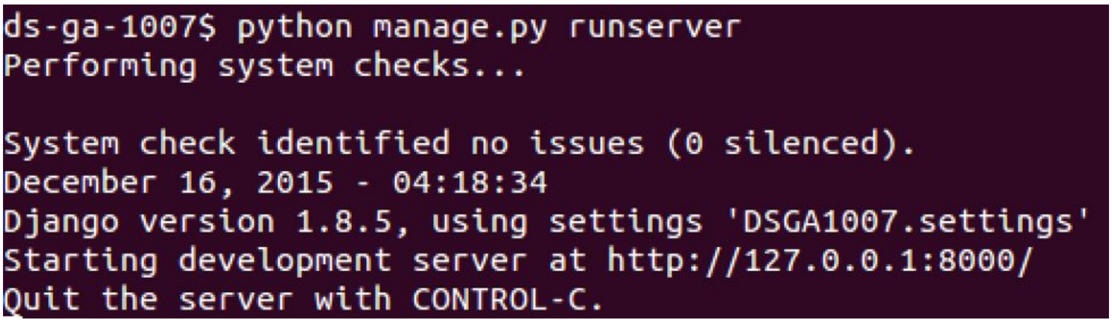


```
ds-ga-1007$ python manage.py makemigrations
```



```
ds-ga-1007$ python manage.py migrate
```

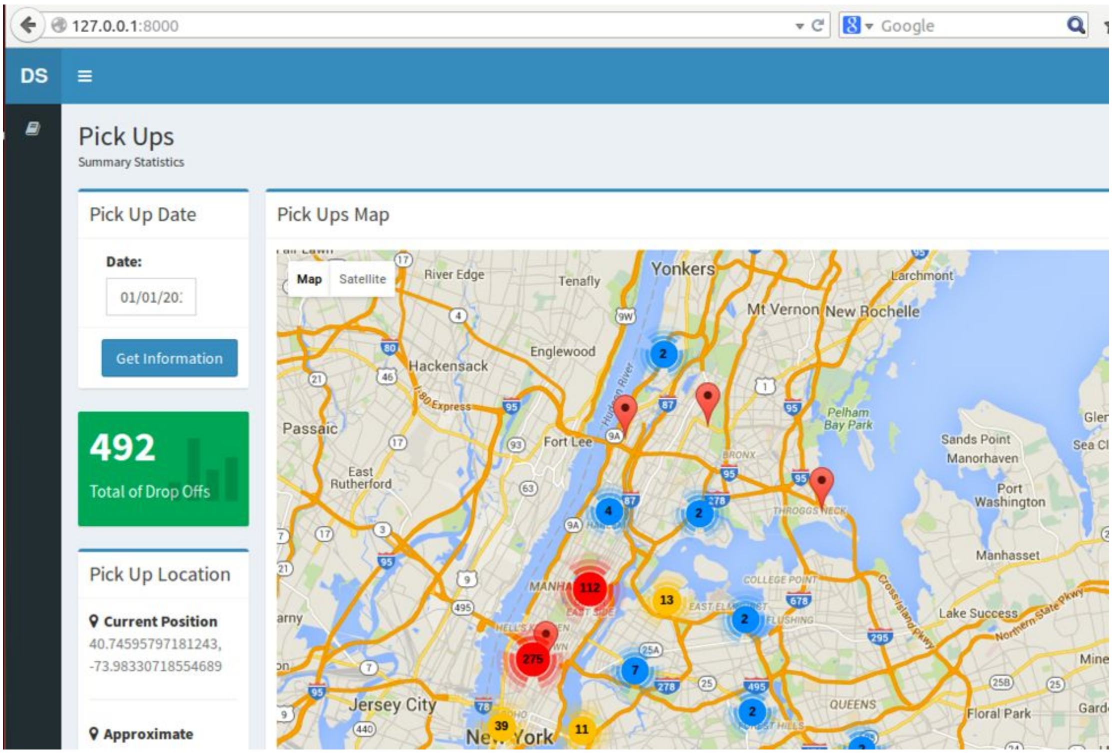
After the two lines above run successfully, type 'python manage.py runserver' in the terminal again. Then you will see something like this:



```
ds-ga-1007$ python manage.py runserver
Performing system checks...

System check identified no issues (0 silenced).
December 16, 2015 - 04:18:34
Django version 1.8.5, using settings 'DSGA1007.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CONTROL-C.
```

*last step: Open the internet server and type “127.0.0.1:8000” in the address bar. You should see something like this and you can select a date from the Pick Up Date section.



Pick Ups
Summary Statistics

Pick Up Date

Date: 01/01/20:

Get Information

492
Total of Drop Offs

Pick Up Location

Current Position
40.74595797181243, -73.98330718554689

Approximate

Pick Ups Map

WATCHME: YouTube link: https://www.youtube.com/watch?v=WTsgC5fWZ6k

Expected Results

The project is divided into different sections:

- The first section, shows all the destinations given a pick up location.

- The second part is focused on the creation of clusters of the drop off locations, being the goal of this, the identification of the 20 destinations given that pick up location. The third section gives us the distribution of the pick ups throughout the day.
- The fourth section are other summary statistics and plots.

The project also allow to download a PDF report.

It is worth mentioning the following:
The project uses the 2015 Yellow Taxi Trip Data. Available at:
(https://data.cityofnewyork.us/view/ba8s-jw6u)
Which contains around 77 million rows. We decided to use a dataset of the first 1 million rows. Unfortunately this only gives us information of 5 days. (2015-01-01, 2015-01-02, 2015-01-08, 2015-01-09, 2015-01-10).
We decided to put the dataset on a MySQL database, make a query and then transform it to a dataframe, so it could be easier to handle the information.

The clusters are represented by circles (which radius is defined by the distance between the centroid and the farthest point in the cluster). This representation could not be the optimal due to the overlapping of these circles, which could cause confusion in the user.