Steps to run code for DIA Final Project:

There are two executables in the project and those needs to be run in the following order

1. Jupyter Notebook (**House_rental.ipynb**)
2. Hadoop Application (**geolocation_distance_calculator.jar**)

Steps to run Jupyter Notebook:

1. Open "Anaconda Command Prompt"
2. Goto the project folder
3. Run "***conda env create -f environment.yml***" to create the conda environment. The name of the environment is "***DIA***"
4. Next activate the environment with "***conda activate DIA***"

Steps to run the Hadoop Application:

1. Ensure "**PostgreSQL**" database is already installed on the system and the tables are created using the code in **Jupyter Notebook**. Please ensure "**listings**" table has all the records before running the Hadoop application. Appropriate code is written in the notebook to insert the cleaned records for all the Airbnb listings.
2. Setup Hadoop in the System
3. Fetch the PostgreSQL JDBC driver from the "***hadoop/dependencies***" in the project folder.
4. Format the namenode using "***hadoop namenode -format***" command
5. Start the Hadoop Framework using "***start_all.cmd***"
6. Type "***hadoop classpath***" to see all the paths in the classpath. Put the "***JDBC Driver***" in one of the file paths already added to the class path to successfully run the application.
7. Place the Hadoop application jar in folder of your choice.
8. After starting Hadoop when all the nodes are ready, run the following command "***hadoop jar <jar file path> <postgresql server username> <postgresql password>***". No need to provide any input/output path. All the records will be retrieved from PostgreSQL db (**ireland_airbnb**) and transformed records will be stored there as well.
9. After the completion of the tasks stop Hadoop server with "***stop_all.cmd***"