HOUSING DATA FOR SF

OVERVIEW:

The dataset to be used was acquired by Controller's office to look at the impact of inclusionary housing requirements on housing development. For this, the collected regression data for Inclusionary Housing Simulation Model in collaboration with Blue Sky Consulting Group, Century Urban LLC, and Street Level Advisors. This dataset has been acquired from SF OpenData which is clearinghouse for data published by city and county of SF. Following are details about the dataset:

1. Source for data: Click here
2. Source for data attributes/columns description: Click here
3. Number of attributes/columns: 106
4. Number of data records: ~2315k
5. Size: 880 MB

INTEREST:

The Controller's Office have certain inclusionary and exclusionary policies for housing requirements. Due to changing economy and various factors, the controller's office enforces
Inclusionary zoning/ Inclusionary Housing policies that links the production of affordable housing to the production of market-rate housing. IZ policies either require or encourage new residential developments to make a certain percentage of the housing units affordable to low or moderate income residents. This project aims at studying the impact of the policies by analysing the regression dataset available. Following are certain patterns I aim to find out.
1. How is the pricing affected? What impact does it have on the stock market?
2. How is the land use pattern changing if any? Are there any specific areas of SF that are being affected regarding commercial or residential use? If so what are the other factors that are causing it?

METHODOLOGY:

The dataset consisted of around ~2315k records. The data set was downloaded from the above mentioned site, filtering by year and obtained 15 files of each year from 2001 to 2015.
Following preprocessing steps were done:

1. python create_parameter_list.py
    Generates file parameter_boolean:
                                    parameter if True is unique for all rows
                                                                else constant

2. python filter_columns.py
   Iterates through the dataset, separates repetitive and unique data (for each row) data for each year
   Common.csv contains constant data of all years ~1kB
                      (24 columns, 15 rows)
   Data_*.csv contains data for the respective year (Boolean representations converted to strings) ~15MB
                      (13 columns, 154342 rows)

3. python create_regionwise_data.py
        iterates through Data_*.csv (* indicates 2001 to 2015)
        Creates area wise files
         (few of the fields that were missing was replaced with the most frequent data elements. Note that: there were a very less number of fields with missing data ~NOT TO BE CONFUSED WITH 'MISSING' VALUE FOR ONE OF THE VARIABLES.)

End Result: 36 files corresponding to each region was obtained. Common.csv containing common indices was obtained.

4. For plotting the Map:
 sf.geojson file was downloaded from [here](#)

 Two regionwise changes were made in the geojson file (on the online tool at geojson.io)
 : removed glen park merged with diamond heights
   Created mission bay from soma region

Number of houses in each area was calculated with help of a script and added as an attribute to geojson file.
(using pygeoj)
Script name: edit_geojson.py

5. Python create_stats.py was used to create simple statistical data row-wise, and analyze the same, select the visualizations accordingly.

6. Project.html contains all of the webpage code.
        Software libs used: dc.js/css, d3.js/css, leaflet.js/css , bootstrap.css

STORY:

Inclusionary housing policies affect the properties. Over the range of 15 years, how each of the properties of each neighbor were affected in terms of different attributes such as
    1. Zoning (Commercial, Residential Use)
    2. Land Use (Industrial, Medical)
    3. Properties having protected Status
    4. Area of property and the actual land use
    5. The percentage of development that can be done.

ANALYSIS:

Initial Plan:
1. Depending upon the kind of land use, the pattern of the indices affected by the same can be concluded.
2. Correlation between indices of stock markets* and price indices*.

Final Analysis:
The indices mentioned above were common for all of the records for a year.
Neighborhood wise, which has largest number land use for industrial use in a particular region.
The land use changes occurring in a particular neighborhood. Same questions were answered in case of the zoning data as well.

WEBPAGE LINK:
edlab.cs.umass.edu/~kchauhan/

ADDITIONAL NOTE:
Following are the variable descriptions:

Landuse:

Missing   Missing
CIE   Cultural, Institutional, Educational
IND   Industrial
MED   Medical
MIPS   Management, Information, Professional Services
Mixed   mixed-use without residential
OpenSpace   OpenSpace
PaperLot   PaperLot
Public   Public
Resid   Residential
RETAIL   (Retail/Entertainment)
ROW   Right-of-way
VACANT   Vacant
VISITOR   Visitor/ Hotel, Visitor services

Zoning:

Commercial   Commercial
DTR   Downtown Residential
Missing   Missing
MixedUse   Mixed Use
Open   Open Space
PDR   Production, Distribution, Repair
Public   Public
RC3   Residential-Commercial, medium density
RC4   Residential-Commercial, high density
Redev   Redevelopment plan areas
Res_Low   Residential (low-density, up to 3-family)
RM1   Residential-Mixed, low-density
RM2   Residential-Mixed, moderate-density
RM3   Residential-Mixed, medium-density
RM4   Residential-Mixed, high-density
RSD   SOMA, residential service
RTO   Residential Transit-Oriented district(s)
SLI   Service/Arts/Light Industrial
SLR   SOMA: Service/Light Industrial/Residential
SSO   SOMA: Service/Secondary/Office