# DATATHON

*FACTORS INFLUENCING HOUSE PRICE*

# PROBLEM STATEMENT

- Given that a ton of open data available, finding the most relevant features that drive the house prices.
  - We tried to make use of diverse data to find interesting features that might drive house prices
  - Interesting features and diverse datasets are explained in the next slide
  - The complexity of this problem lies in cleaning and aggregating diverse data, and integrating it in the existing Enigma's dataset.
  - The second level of complexity lies in understanding and extracting relevant and important features from open data

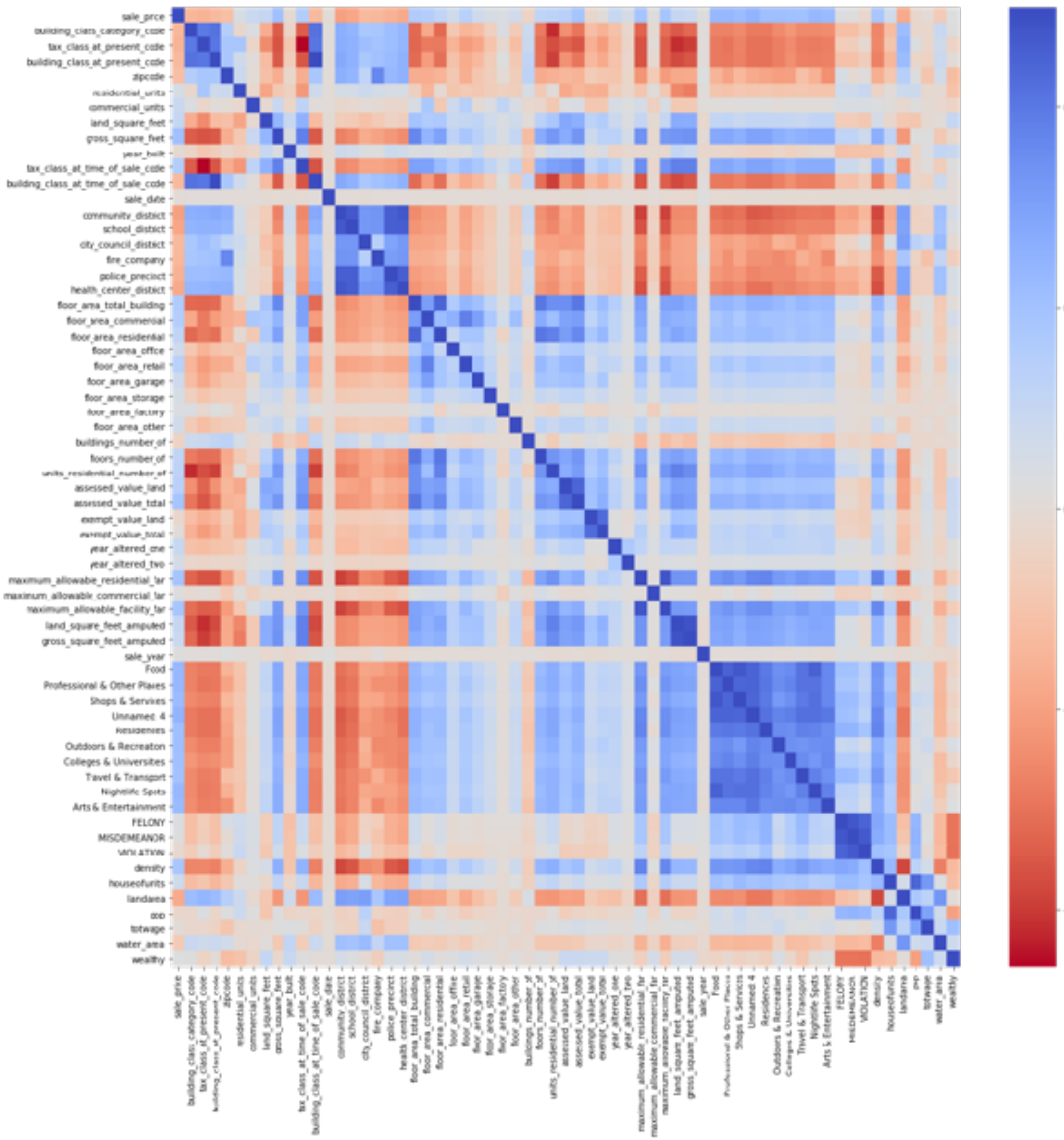
# WHAT MATTERS

- Crime Rate
- Accessibility
- Popularity
- Type of population
- Fun things to do around
- Educational Facilities around
- Of course, properties of the house

# DATA SETS

- **Enigma NYC Property Sale:**
  - Accessed value land, Floor area residential, Floor area total building, Year built, Residential Units, altering history
  - Gives information about  the particular house and its properties.
- **FourSqure Data:**
  - Gives following venue information about Zipcode: Professional places, shops&services, Residence, Outdoor&Recreation, College & University, Travel&Transport, Night life&Sports, Arts & Entertainment.
  - Gives idea about how engaging one area is
- **USZipcode API:**
  - Converts latitude and longitude information to zip code
  - Gives other factors like : Number of House Units, Density, Land Area, Water Area, Population, Total wage, Wealthy
  - Gives intuition about quality of life and population and demographic information
- **NYU Yellow Taxi:**
  - Number of pickups and drop offs from particular zip code
  - Gives information about how accessible area is.
- **NYC Crime Data**
  - Gives following information based on Zipcodes: Felony, Misbehavior, Violation, Misdemeanor
  - Gives intuition about safety of an area

# CORRELATION AMONG FEATURES



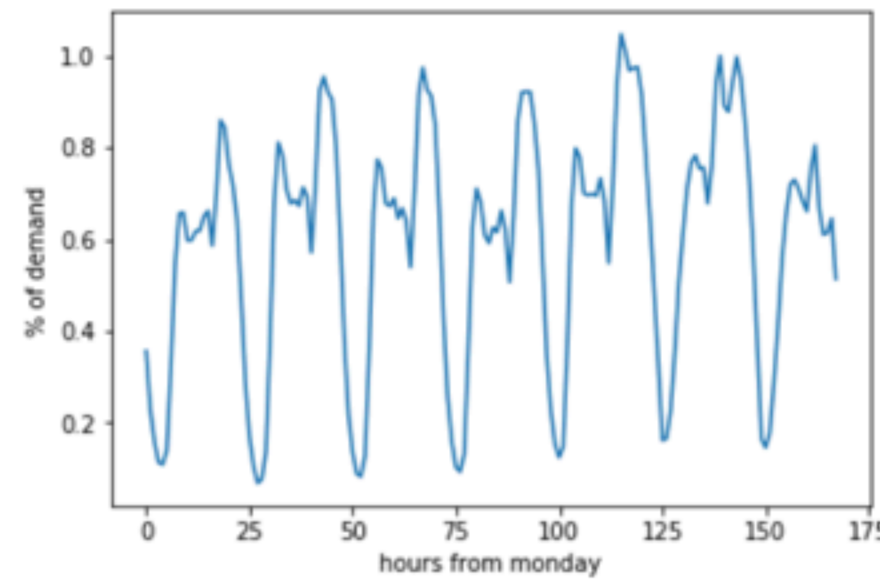

- *Shops& Serivices*
- *Residence*
- *Out door & recreation*
- *Food*
- *Travel & Transport*
- *College & Recreation*
- *Night Life*
- *Art & Entertainment*
- *Density*
- *Land Area*
- *Year Built*

# TAXI DATA

- For taxi data, only one months data was made use
- Each month's data is about 2 GB and with time constraint, its impossible to get required values for each months
- The distribution of number of taxis taken over 2 months didn't seem to differ a lot.

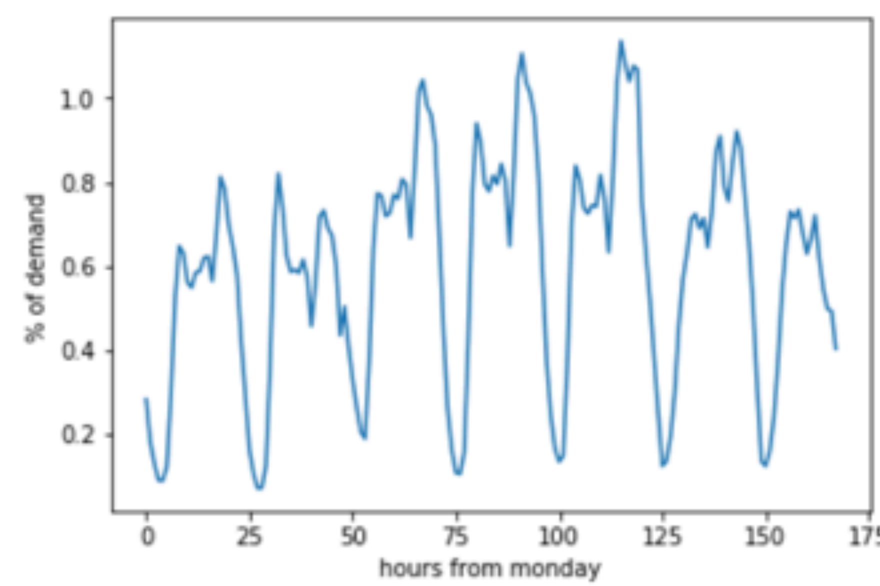
In [7]:

```
complete_function(dffull_feb,1,1,0,0,'hour')
```

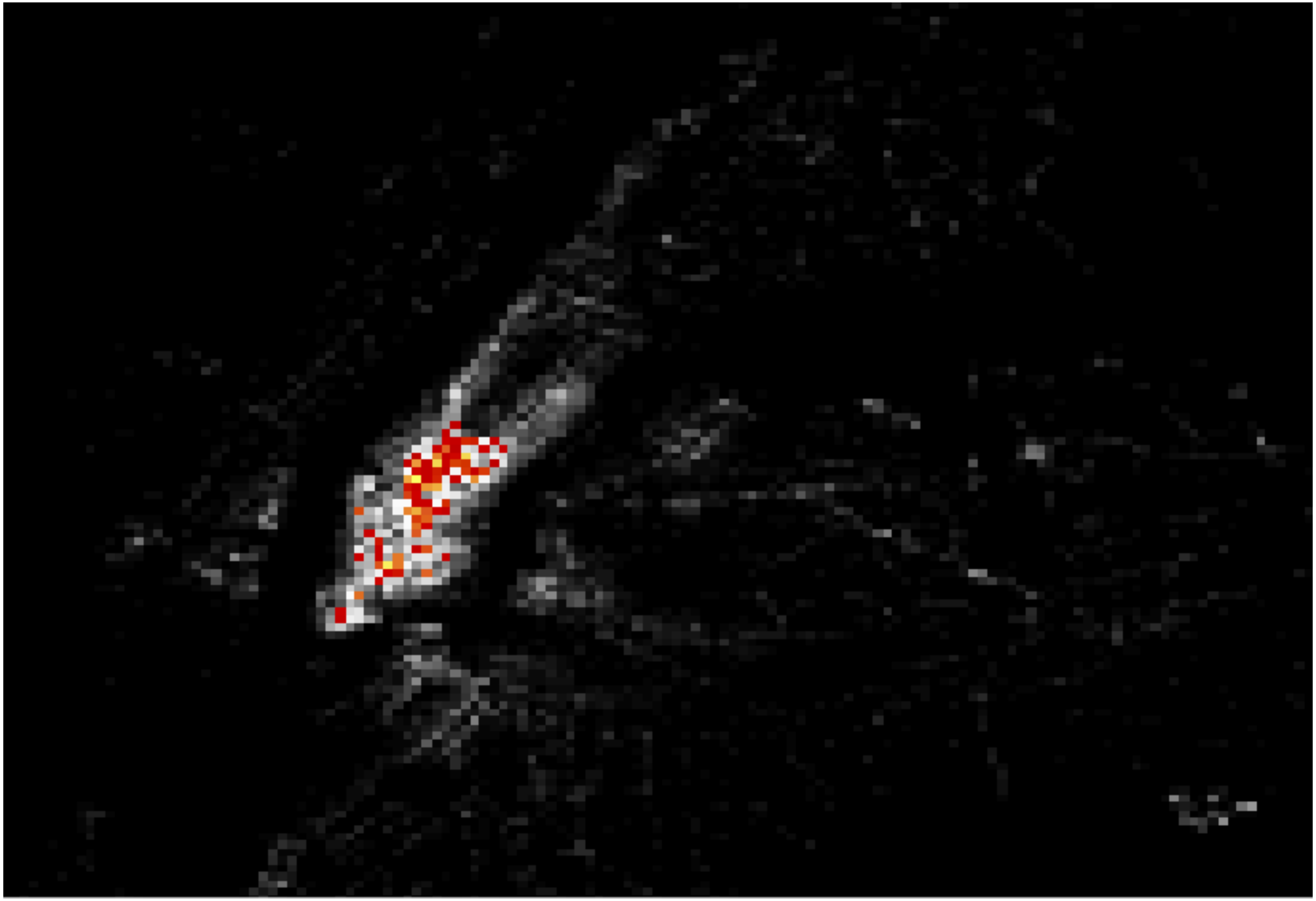



In [9]:

```
complete_function(dffull_jan,1,1,0,0,'hour')
```



# NYC TAXI



# FOUR SQUARE

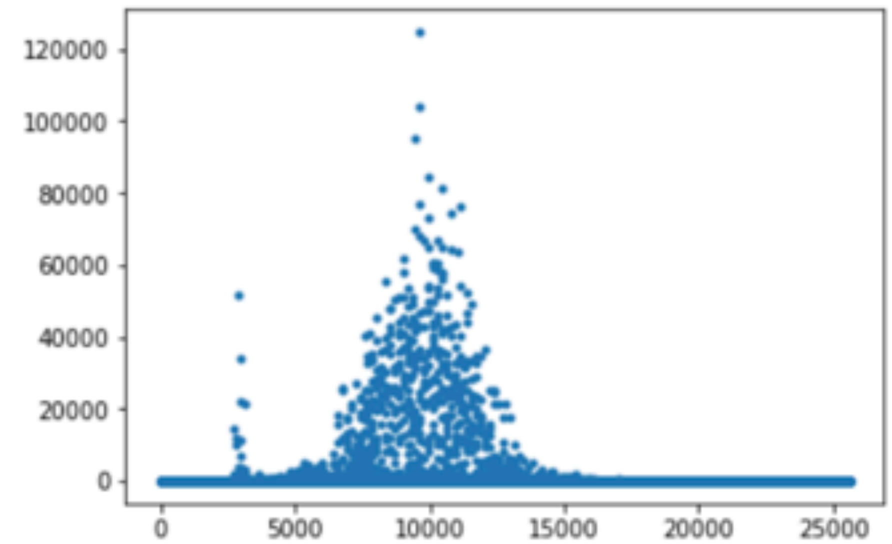
*Number of taxis taken and*

*Number of foursquare venues*

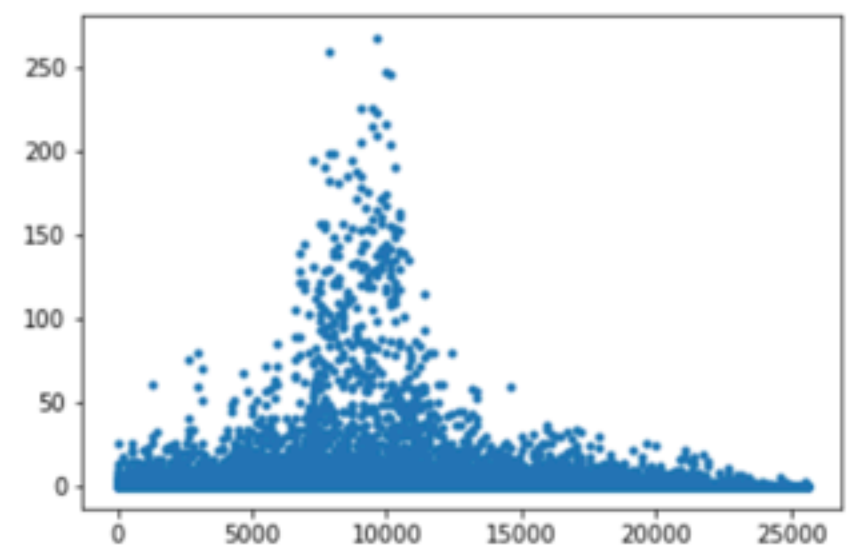
*are correlated (0.82)*

**Plot of Number of venues and Taxi taken -Grid wise**

In [314]:
```
plt.plot(np.arange(len(jan_count[0])),jan_count[0],'.')
plt.show()
```

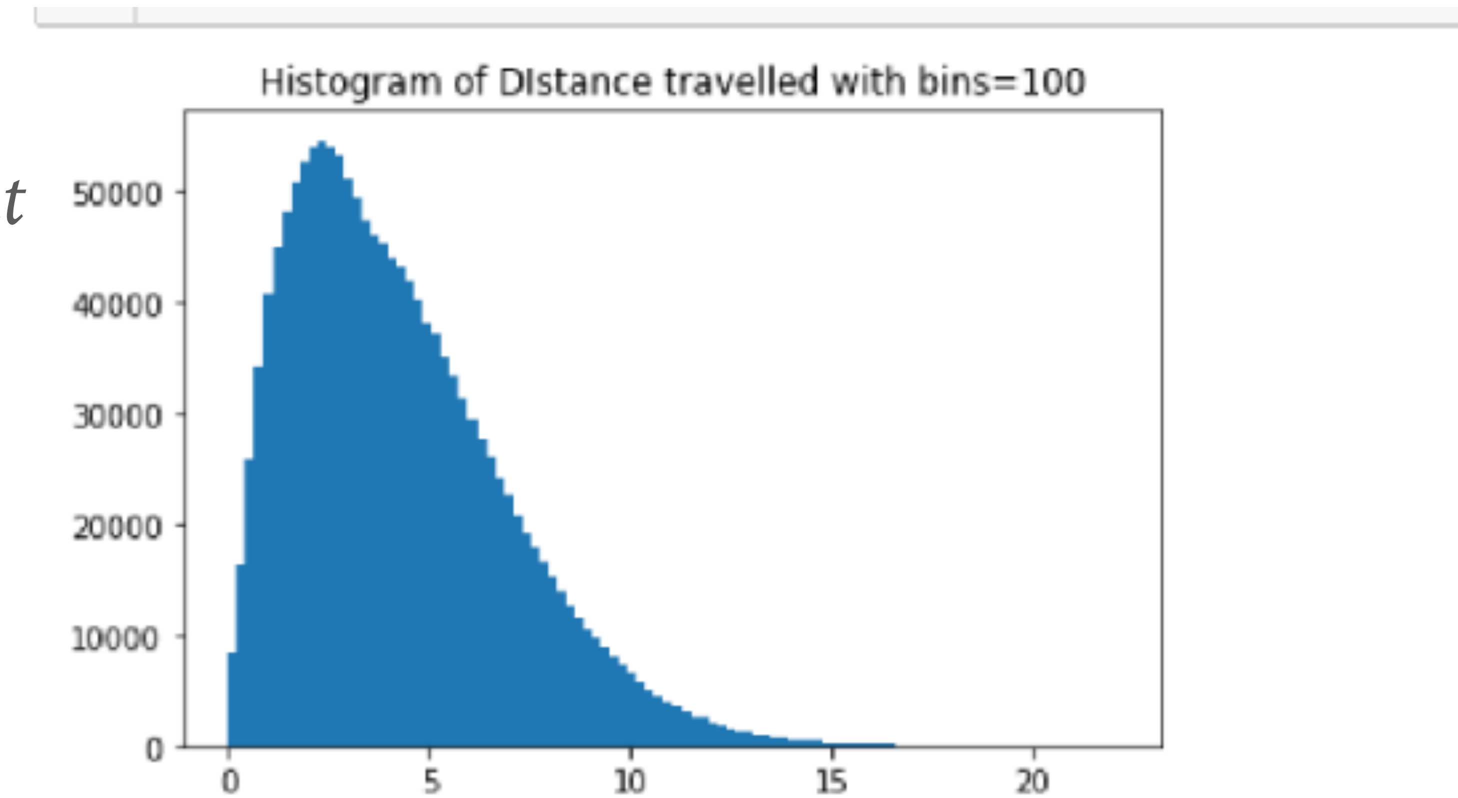



In [199]:
```
zx=four_cell.reshape((1,-1))
plt.plot(zx[0],'.')
plt.show()
```



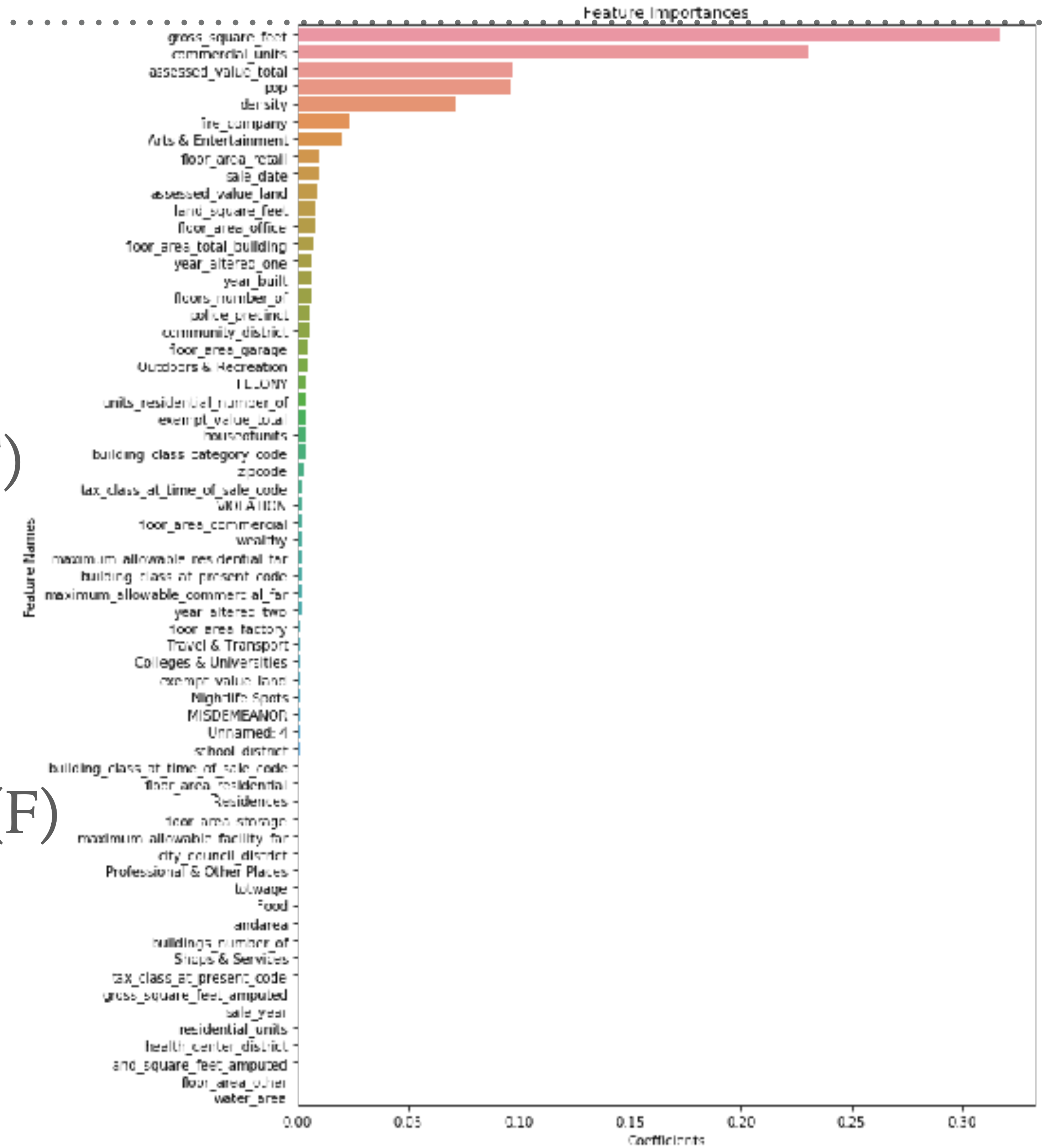
# WHY FOURSQUARE VENUE AFFECT HOUSE PRICE

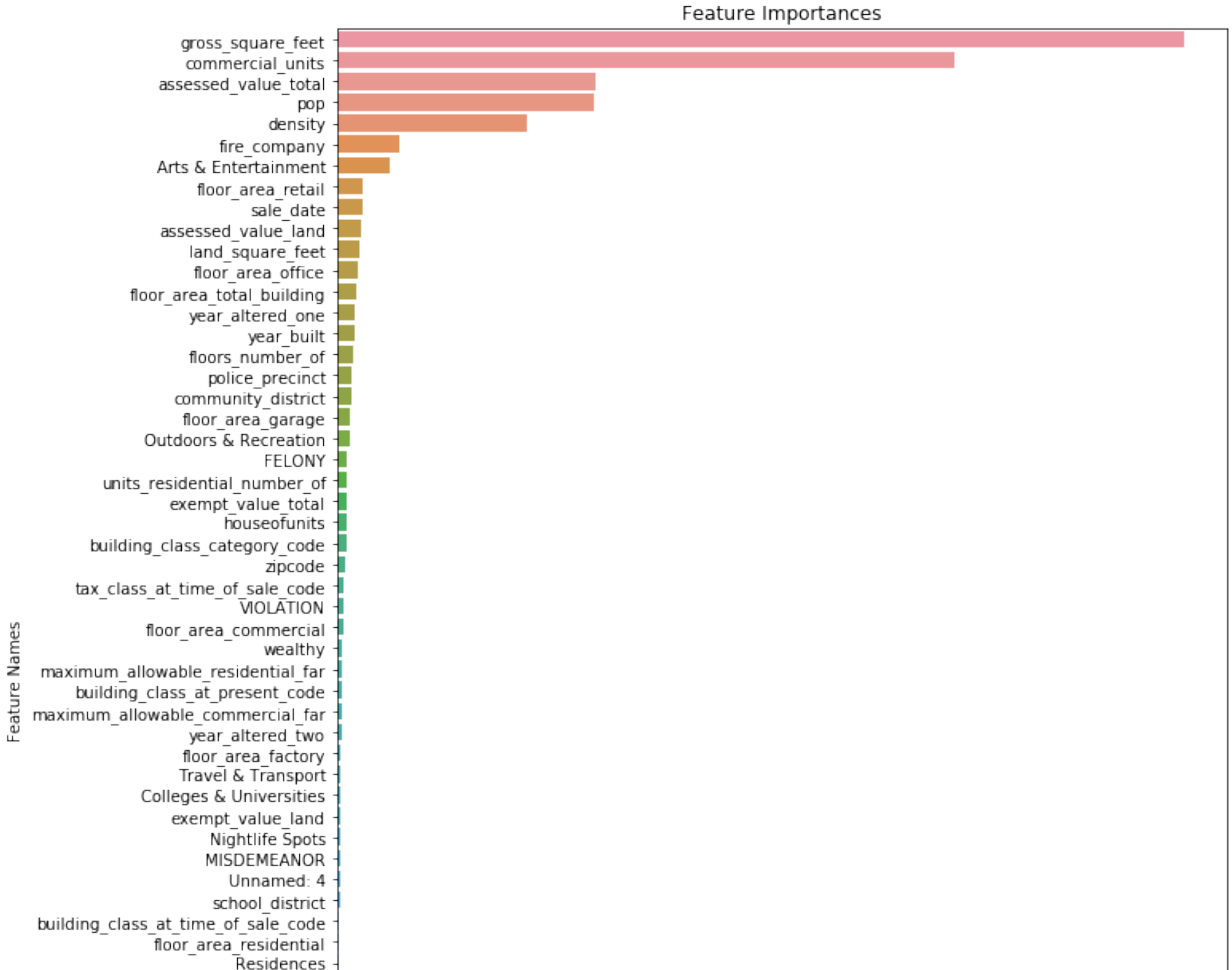
*From Taxi trips, its evident that People usually prefer traveling at max 5 km.*



# FEATURE IMPORTANCE

- ➤ Gross square feet (E)
- ➤ Commercial Units(E)
- ➤ Density(Z)
- ➤ Art & Entertainment (F)
- ➤ Police precinct(E)
- ➤ Felony (C)
- ➤ Outdoor & Recreation (F)
- ➤ Land Square feet (E)





Feature Importances
Feature Names
gross_square_feet
commercial_units
assessed_value_total
pop
density
fire_company
Arts & Entertainment
floor_area_retail
sale_date
assessed_value_land
land_square_feet
floor_area_office
floor_area_total_building
year_altered_one
year_built
floors_number_of
police_precinct
community_district
floor_area_garage
Outdoors & Recreation
FELONY
units_residential_number_of
exempt_value_total
houseofunits
building_class_category_code
zipcode
tax_class_at_time_of_sale_code
VIOLATION
floor_area_commercial
wealthy
maximum_allowable_residential_far
building_class_at_present_code
maximum_allowable_commercial_far
year_altered_two
floor_area_factory
Travel & Transport
Colleges & Universities
exempt_value_land
Nightlife Spots
MISDEMEANOR
Unnamed: 4
school_district
building_class_at_time_of_sale_code
floor_area_residential
Residences

# INSIGHTS FROM FEATURE IMPORTANCE

- As expected, diverse datasets did give some idea about the different attributes that correlates with the price of a house, which we could also interpret as making a place for desirable to live.
- Intuitively Gross Square Feet and Commercial Units are important trivially.
- A high density naturally suggest high demand for living in the area.
- Felony and police precinct can be used as a proxy for how safe the neighbourhood.

# CONCLUSION

- We focussed our efforts on finding the attributes that primarily drive the price of a property. As as natural next step, we would like to gather more data and built a model to predict prices primarily based on the features which we found to be important.

# THANKS