2021F-T1 AISC1006 – Step Presentation 01 (M07 Group 1)

Student Name(s) ↓	Student IDs ↓
Shivam Patel	500201461
Jaydeep Bhalala	500198056
Pratik Domadiya	500199494

2021F-T1 AISC1006 - Step Presentation (Step 1) 01 (M07 Group 1)

Dataset Preparation

Name : Pratik Domadiya
Loyalist ID : 500199494
pratikdomadiya@loyalistcollege.com

2021F-T1 AISC1006 - Step Presentation (Step 1) 01 (M07 Group 1)

Data Cleaning and Modelling

Name: Jaydeep Bhalala

Student ID-500198056

jaydeepbhalala@loyalistcollege.com

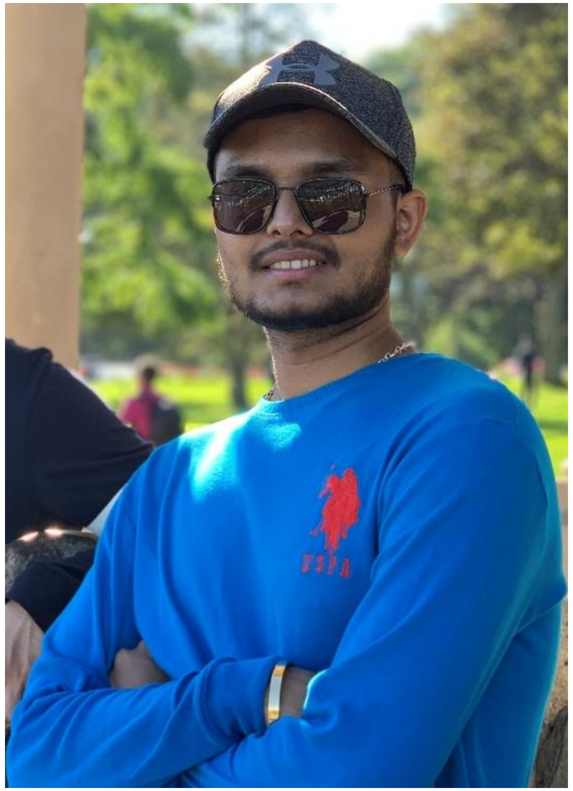
2021F-T1 AISC1006 - Step Presentation (Step 1) 01 (M07 Group 1)

Data Visualization

Name: Shivam Patel

Student ID-500201461

shivampatel@loyalistcollege.com



ROAD MAP

1. Approach and source code
2. Visualization & Presentation
3. Analysis & Experience

1. Approach and source code

Data set preparation

Data Cleaning

Visualization & Analysis

Model Preparation

Final Result Evaluation

Data Preparation

Concept

- We have created our own dataset.
- Web Scraping

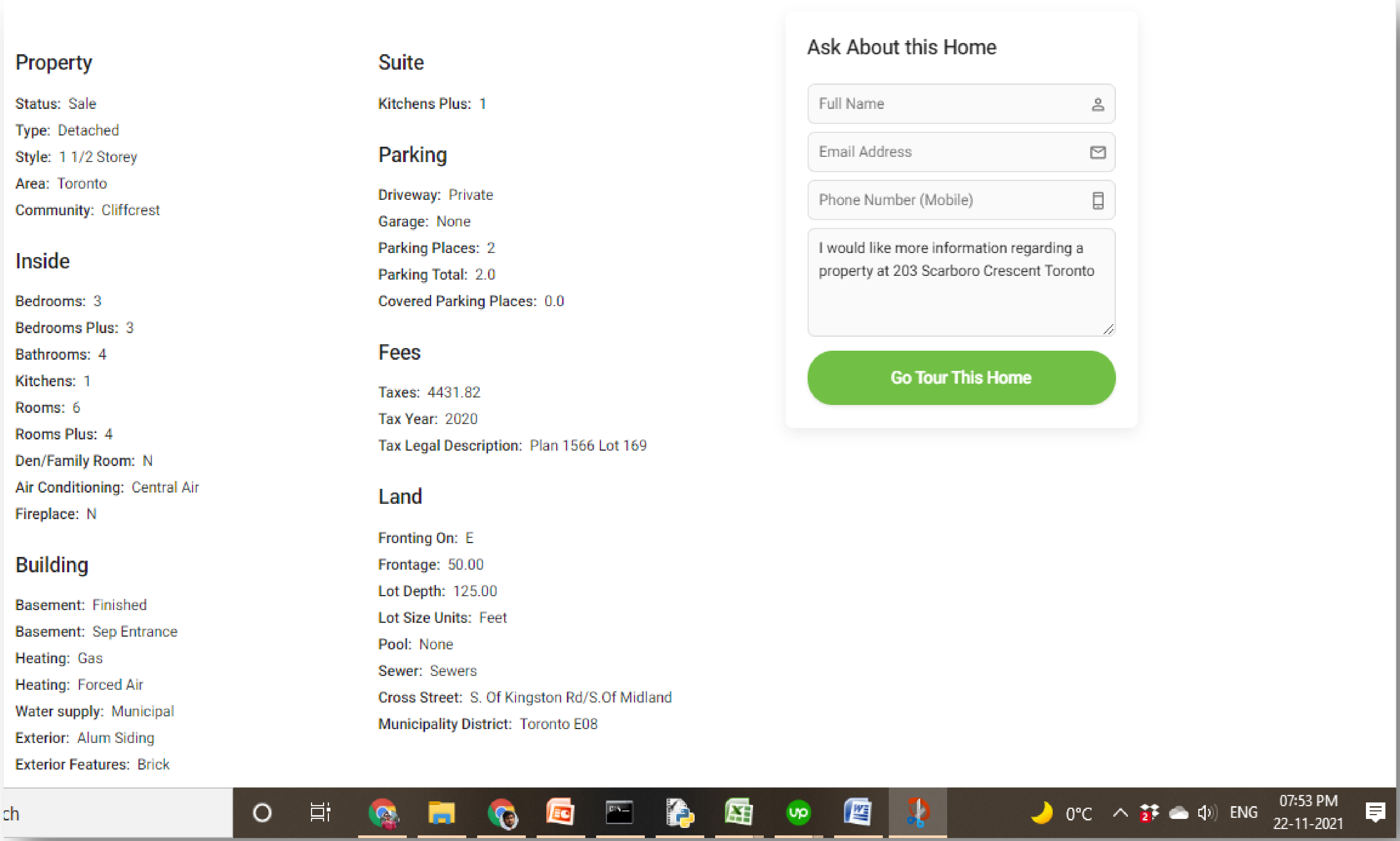
Technology

- Python
- Beautiful soup 4
- Requests library
- Selenium web driver

Data Source

- Zolo Toronto website

sample data on zolo website



Scrapping Result

- Output stored in excel file. Please have a look on features.

TYPE
WALK SCORE
SIZE (SQ FT)
BEDROOMS
BATHROOMS
KITCHENES
ROOMS
DEN / FAMILY ROOM
PATIO TERRACE
ENSUITE LAUNDRY
AIR CONDITIONING
FIREPLACE
STORIES
PARKING TOTAL
MAINTAINANCE (PER ANNUM)
TAXES
BASEMENT
HOUSE_PRICE

SAMPLE DATASET

A1499 | f_x | Commercial/Retail

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Type	Walk Score	Age	Listed By	Size (sq ft)	Bedrooms	Bathrooms	Kitchens	Rooms	Area	/Family Ro	atio Terrac	suite Laund	Conditioni	Fireplace	Stories	arking Tota	laintenanc	Taxes	Community	Basement	house_price
2	Condo Townhouse	83	No Data	Royal Lep	1400-1599	3	3	1	8	Toronto	N	Terr	Y	Central Ai	N	2	1.0	446.40	3250.59	Dovercou	None	999900
3	Condo Apt	37	No Data	Royal Lep	900-999	2	1	1	6	Toronto	N	Open	Y	Central Ai	N	15	1.0	697.42	1255.60	Flemingd	None	525000
4	Condo Apt	97	No Data	Homelife	600-699	1	1	1	5	Toronto	N	Open	Y	Central Ai	N	19	0.0	435.14	2480.71	Church-Yc	None	499999
5	Condo Townhouse	30	No Data	Re/max N	1000-1199	3	2	1	5	Toronto	N	Terr	Y	None	Y	1	2.0	649.93	1438.29	Elms-Old	Crawl Spa	529800
6	Detached	37	No Data	First Clas	3500-5000	4	6	1	10	Toronto	Y	NO	N	Central Ai	Y	0	4.0	0.0	15617.00	St. Andrev	Walk-Up	4280000
7	Condo Apt	70	No Data	Right At H	500-599	1	1	1	4	Toronto	N	Open	Y	Central Ai	N	20	1.0	512.11	1619.18	Islington-	None	535000
8	Condo Apt	79	0-5	Living Rea	700-799	1	2	1	5	Toronto	N	Open	Y	Central Ai	N	9	1.0	553.52	2309.63	Waterfror	None	778000
9	Detached	77	No Data	Right At Home Real		3	2	1	9	Toronto	N	NO	N	Central Ai	N	0	2.0	0.0	3531.65	Keelesda	Walk-Up	998888
10	Condo Apt	38	No Data	Century 2	600-699	1	1	1	4	Toronto	N	Open	Y	None	N	10	1.0	657.15	212.20	Black Cre	None	149900
11	Condo Apt	48	No Data	Re/max R	1000-1199	2	2	1	5	Toronto	N	None	Y	Central Ai	N	14	1.0	721.98	1115.45	Mount Ol	None	579000
12	Detached	82	No Data	Tailored Realty Inc.		2	2	2	4	Toronto	N	NO	N	None	N	0	2.0	0.0	3298.38	The Beacl	Sep Entra	1399999
13	Att/Row/Twnhouse	91	51-99	Harvey Ka	1500-2000	3	3	2	7	Toronto	Y	NO	N	Central Ai	Y	0	0.0	0.0	4998.00	Kensingtc	Full	1279000
14	Condo Apt	98	No Data	Right At H	500-599	1	1	1	4	Toronto	N	Open	Y	Central Ai	N	16	1.0	417.12	2529.59	Waterfror	None	599000
15	Condo Townhouse	72	No Data	Bay Stree	1400-1599	4	4	1	7	Toronto	N	None	N	Central Ai	Y	1	2.0	398.00	2419.00	Agincourt	Finished	699900
16	Detached	40	No Data	Exp Realt	2000-2500	4	4	1	8	Toronto	Y	NO	N	Central Ai	Y	0	4.0	0.0	4668.12	Steeles	Finished	1288888
17	Detached	31	No Data	Homecomfort Realt		5	6	1	10	Toronto	Y	NO	N	Central Ai	Y	0	4.0	0.0	4264.87	L'Amorea	Finished	999000
18	Condo Apt	90	No Data	Re/max R	600-699	1	1	1	5	Toronto	N	Open	Y	Central Ai	N	18	1.0	420.22	2348.07	Waterfror	None	724900
19	Detached	57	No Data	Ipro Real	1100-1500	3	2	1	6	Toronto	N	NO	N	Central Ai	N	0	3.0	0.0	3372.80	Dorset Pa	Finished	899000
20	Condo Apt	80	0-5	Royal Lep	700-799	2	2	1	5	Toronto	N	Open	Y	Central Ai	N	6	1.0	689.58	3116.16	Waterfror	None	829900
21	Condo Apt	18	No Data	Right At H	800-899	2	1	1	6	Toronto	Y	Open	Y	Central Ai	N	24	1.0	485.48	1655.85	West Hur	None	569900
22	Detached	91	100+	Sutton Gr	2500-3000	5	4	2	11	Toronto	Y	NO	N	Central Ai	Y	0	2.0	0.0	7008.32	Moss Parl	Finished	1799000
23	Condo Apt	97	0-5	Homelife	600-699	1	1	1	4	Toronto	N	Open	Y	Central Ai	N	4	0.0	414.33	2767.89	University	None	729900
24	Condo Apt	97	New	Re/max H	500-599	1	1	1	4	Toronto	N	Open	Y	Central Ai	N	52	0.0	420.00	238.06	Bay Stree	None	729800
25	Condo Townhouse	66	No Data	Century 2	1500-2000	3	4	1	7	Toronto	Y	None	Y	Central Ai	N	1	2.0	152.00	3200.00	Dorset Pa	W/O	799000
26	Semi-Detached	76	No Data	Royal Lepage Terre		3	1	1	6	Toronto	N	NO	N	Central Ai	N	0	0.0	0.0	4307.64	Greenwo	Unfinishe	899999
27	Condo Apt	68	No Data	Bay Stree	1000-1199	2	2	1	6	Toronto	N	Open	Y	Central Ai	N	2	1.0	708.26	2810.60	Bayview \	None	620000
28	Condo Apt	82	16-30	Re/max C	700-799	1	1	1	4	Toronto	N	Open	Y	Central Ai	N	8	1.0	776.01	3103.95	Waterfror	None	849999

2. Visualization & Analysis

Data Cleaning

- Purpose of the Data cleaning
  - Data cleaning is the process of removing data that is wrong, inaccurate, incomplete, poorly structured, duplicated, or simply unrelated to the dataset's goal.

Cleaning the data

Cleaning of Type Column of the data

```
# checking the unique values available in 'type' column
df['Type'].nunique(),df['Type'].unique()

# Some unique values are merged as they are having the same meaning
# "condoxxxxxx" -> condo
# co-op apt, Apartments, Co-Ownership Apt -> Apartments
# twnhouse -> town house
# triplex, multiplex, duplex, Fourplex -> multiplex
# Single Family,House/Single Family-> Single Family
# Office, Store W/Apt/Office' -> office
# Link, Parking Space, Locker, Parking, Land, Vacant Land -> ohters
# retail, Commercial/Retail -> retail
```

```
(33, array(['Condo Townhouse', 'Condo Apt', 'Detached', 'Att/Row/Twnhouse',
        'Semi-Detached', 'Comm Element Condo', 'Co-Op Apt', 'Investment',
        'Det Condo', 'Vacant Land', 'Apartments', 'Condominium', 'Triplex',
        'Link', 'Multiplex', 'Condo/Apt Unit', 'Other', 'Co-Ownership Apt',
        'Duplex', 'Parking Space', 'Locker', 'Single Family',
        'Commercial/Retail', 'Industrial', 'Land', 'Office',
        'Row / Townhouse', 'Leasehold Condo', 'Fourplex',
        'House/Single Family', 'Retail', 'Store W/Apt/Office', 'Parking'],
       dtype=object))
```

After applying Filters

```
[46] #getting all the unique types of 'Type' column
     df['Type'].unique()
```

```
array(['condo', 'Detached', 'townhouse', 'Semi-Detached', 'Apartments',
       'Investment', 'Other', 'multiplex', 'single_family', 'retail',
       'Industrial', 'Office'], dtype=object)
```

Cleaning of Walk Score Column of the data

- Taking average of walk score by grouping the community.
- Filling null values of walk score based on the mean value of the community

```
#removing the '-' from walk score
temp_df = df.copy()
temp_df['Walk Score'] = pd.to_numeric(temp_df.where(temp_df['Walk Score'] != '-')['
temp_df = temp_df.groupby('Community')[['Walk Score','Bedrooms']].mean().astype('in

for i in range(0,len(df.index)):
  if df.iloc[i]['Walk Score'] == '-':
    df.loc[i, 'Walk Score'] =  temp_df.loc[df.iloc[i]['Community']]['Walk Score']

df.drop(labels=['Community'],axis=1,inplace=True) #We no longer need the 'Community
```

Cleaning of Patio Terrace Column of the data

- Before the cleaning

```
[61] # unique values in Patio Terrace column
     df['Patio Terrace'].unique()

     array(['Terr', 'Open', 'NO', 'None', 'Jlte', 'Encl', 4], dtype=object)
```

- After the cleaning

```
df['Patio Terrace'] = df['Patio Terrace'].apply(process_patio)
df['Patio Terrace'].unique()

array(['Yes', 'No'], dtype=object)
```

Cleaning of Air Conditioning Column of the data

- If present then ‘Y’ else ‘N’

Before the cleaning

```
[54] # unique values in 'Air Conditioning' column
     df['Air Conditioning'].unique()
```

```
array(['Central Air', 'None', 'Window Unit', 'Wall Unit', 'Other',
       'def no data', 'Central Air Conditioning', 'Part', 'Y', 'N'],
      dtype=object)
```

After the cleaning

```
df['Air Conditioning'] = df['Air Conditioning'].apply(process_conditioning)
df['Air Conditioning'].unique()
```

```
array(['Y', 'N'], dtype=object)
```

Cleaning of Stories Column of the data

- Converting every code into numbers

```python
# replacing values with below list
zero = ['Ph','Lph','Th']
one = ['P1','0-1','1&2','A','Low','1st','L15','M']
three = ['3&4']

def process_stories(stroy):
  if stroy in zero:
    return 0
  elif stroy in one:
    return 1
  elif stroy in three:
    return 3
  else:
    return stroy

df['Stories'] = df['Stories'].apply(process_stories)
df['Stories'].unique()
```

```
array([ 2, 15, 19,  1,  0, 20,  9, 10, 14, 16, 18,  6, 24,  4, 52,  8, 17,
       11,  7, 34,  3, 37, 27, 13, 12,  5, 28, 21, 44, 42, 26, 30, 32, 74,
       25, 40, 23, 36, 22, 29, 38, 50, 31, 39, 46, 64, 35, 45, 33, 51, 47,
       54, 48, 65, 59, 41, 53, 56, 49, 79, 58, 69, 60, 57, 43])
```

Cleaning of Basement Column of the data

- If present then it then ‘Y’ else ‘N’

Before the cleaning

```
[74] # unique values in 'Basement' column
     df['Basement'].unique()

array(['None', 'Crawl Space', 'Walk-Up', 'Sep Entrance', 'Full',
       'Finished', 'W/O', 'Unfinished', 'Part Bsmt', 'Part Fin',
       'Fin W/O', 'Apartment', 'Other', 'Half', 'Y', 'No', 'Yes',
       'Partial (Finished)', 'Full (Unfinished)'], dtype=object)
```

After the cleaning

```
df['Basement'] = df['Basement'].apply(process_basement)
df['Basement'].unique()

array(['N', 'Y'], dtype=object)
```

Cleaning of Size Column of the data

- We will use four features('Bedrooms’, 'Bathrooms’, 'Kitchens’, 'Rooms’) to predict the null values of size column by linear regression model.

Null values Before the cleaning

```
[ ] df['Size (sq ft)'].isnull().sum()
752
```

Null values After filling it

```
[188] df['Size (sq ft)'].isnull().sum()
0
```

Null values in the data after cleaning

```
# checking null values if there is any
df.isnull().sum()
```

```
Type                0
Walk Score          0
Size (sq ft)        0
Bedrooms            0
Bathrooms           0
Kitchens            0
Rooms               0
Den/Family Room     0
Patio Terrace       0
Ensuite Laundry     0
Air Conditioning    0
Fireplace           0
Stories             0
Parking Total       0
Maintenance         0
Taxes               0
Basement            0
house_price         0
dtype: int64
```

Data Visualization

Analyze Your Audience

Choose the Right Chart (i.e., Scatterplot, Boxplot, histogram, etc.)

Select a Software Program (we used Jupyter Notebook)

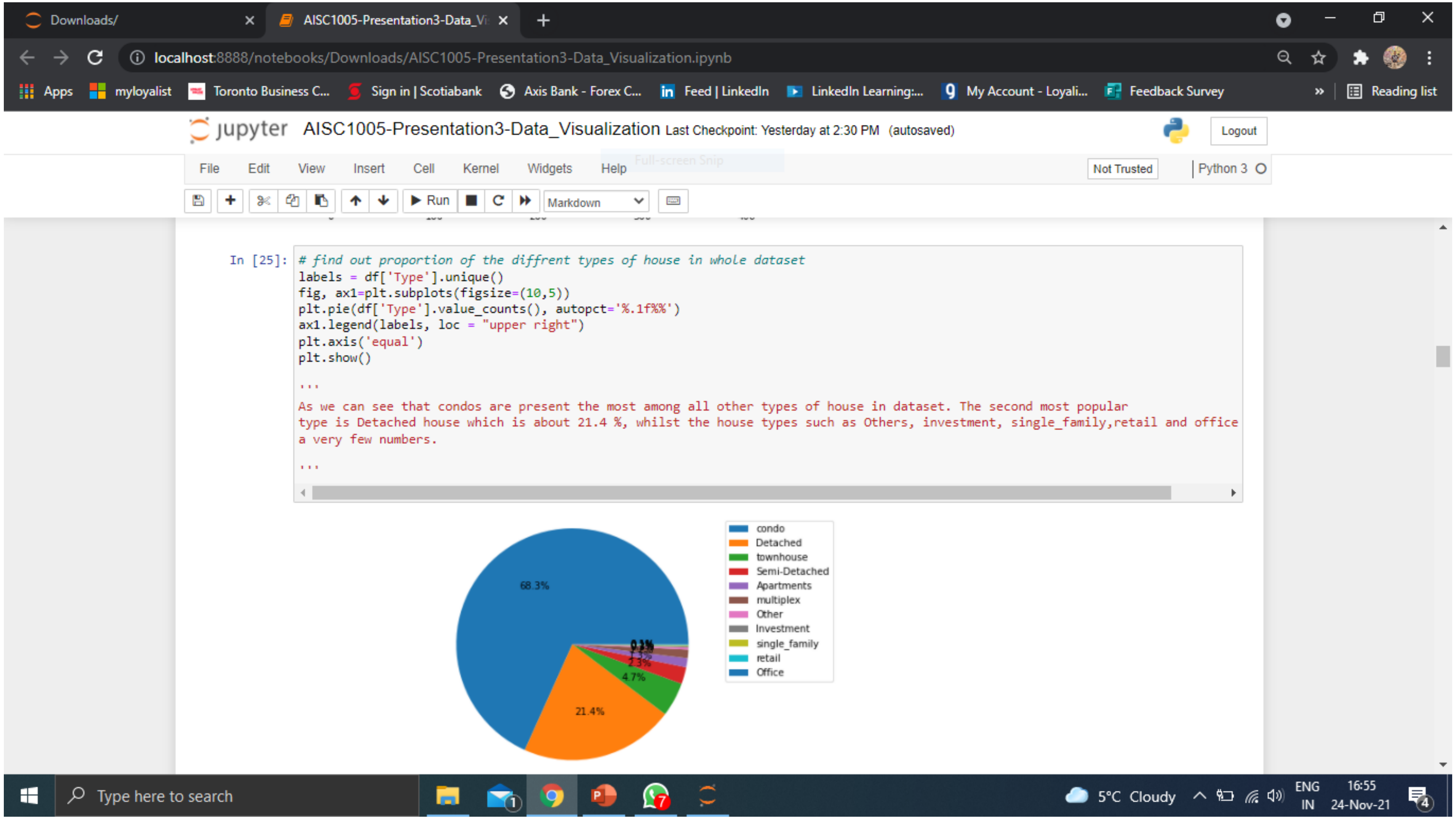
Declutter

Clarify Your Message with Color

Clarify Your Message with Text (i.e., Clear Insights)

Share Your Completed Data Visualization

Sample Coding and Output



- Whole data visualization coding file is submitted with the ppt named AISC1005-Presentation3-Data_Visualization.ipynb

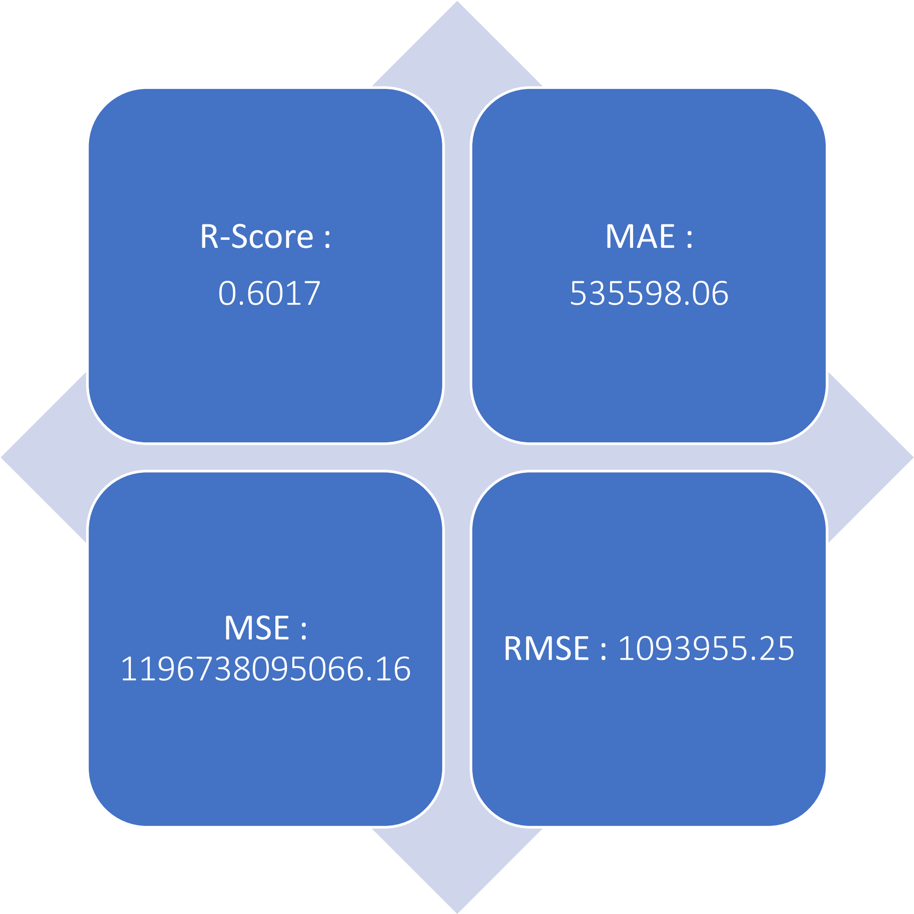
Model Preparation

- We have use 2844 samples for training the linear **regression model** and 948 samples for validating model. For each sample we are providing features that are given below.

```
[22] regression_model.feature_names_in_

array(['Walk Score', 'Size (sq ft)', 'Bedrooms', 'Bathrooms', 'Kitchens',
       'Rooms', 'Den/Family Room', 'Patio Terrace', 'Ensuite Laundry',
       'Air Conditioning', 'Fireplace', 'Stories', 'Parking Total',
       'Maintenance', 'Taxes', 'Basement', 'Apartments', 'Detached',
       'Investment', 'Office', 'Other', 'Semi-Detached', 'condo',
       'multiplex', 'retail', 'single_family', 'townhouse'], dtype=object)
```

Metrics uses to measure Regression model



3. Analysis & Experience

Personal Analysis – Data Preparation

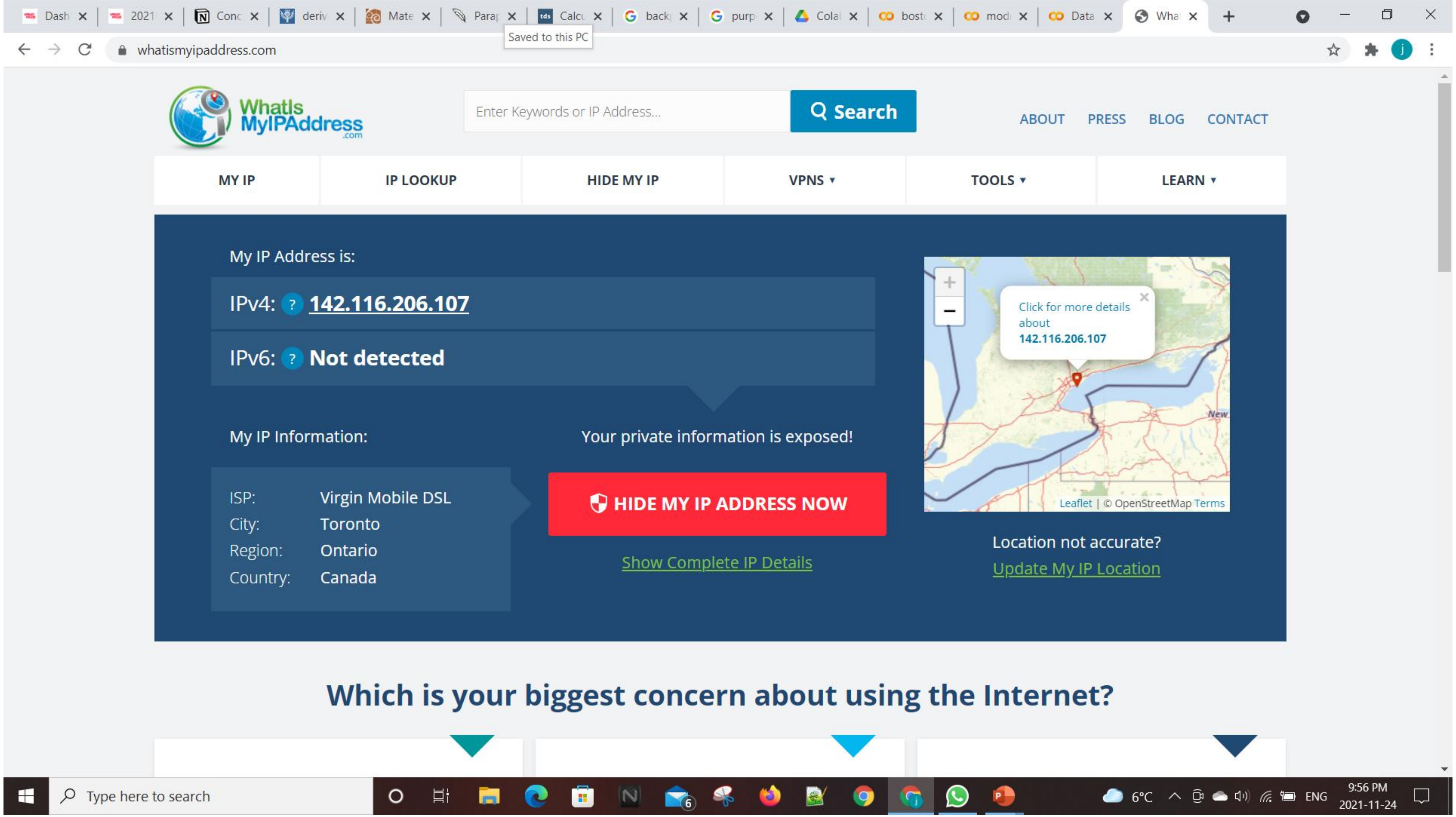
- We have **created our own dataset instead of using just toy dataset** which gives always a high accuracy. By doing so we have learned how web scraping concepts have been used for developing dataset, organizing dataset and how to extract critical features which takes the model accuracy at high level.
- The one thing I have observed that by doing such a project , **data preprocessing takes too much time** when you deal with the completely unknown data. This is because of you don’t know actually what values are stored in each feature columns, how much missing values are existed and so on, so you have to identify missing and unique values are present in each column and then remove unnecessary data values. In our case out of total time taken by the whole project, only 60-70 % taken by the data preparation and preprocessing task.

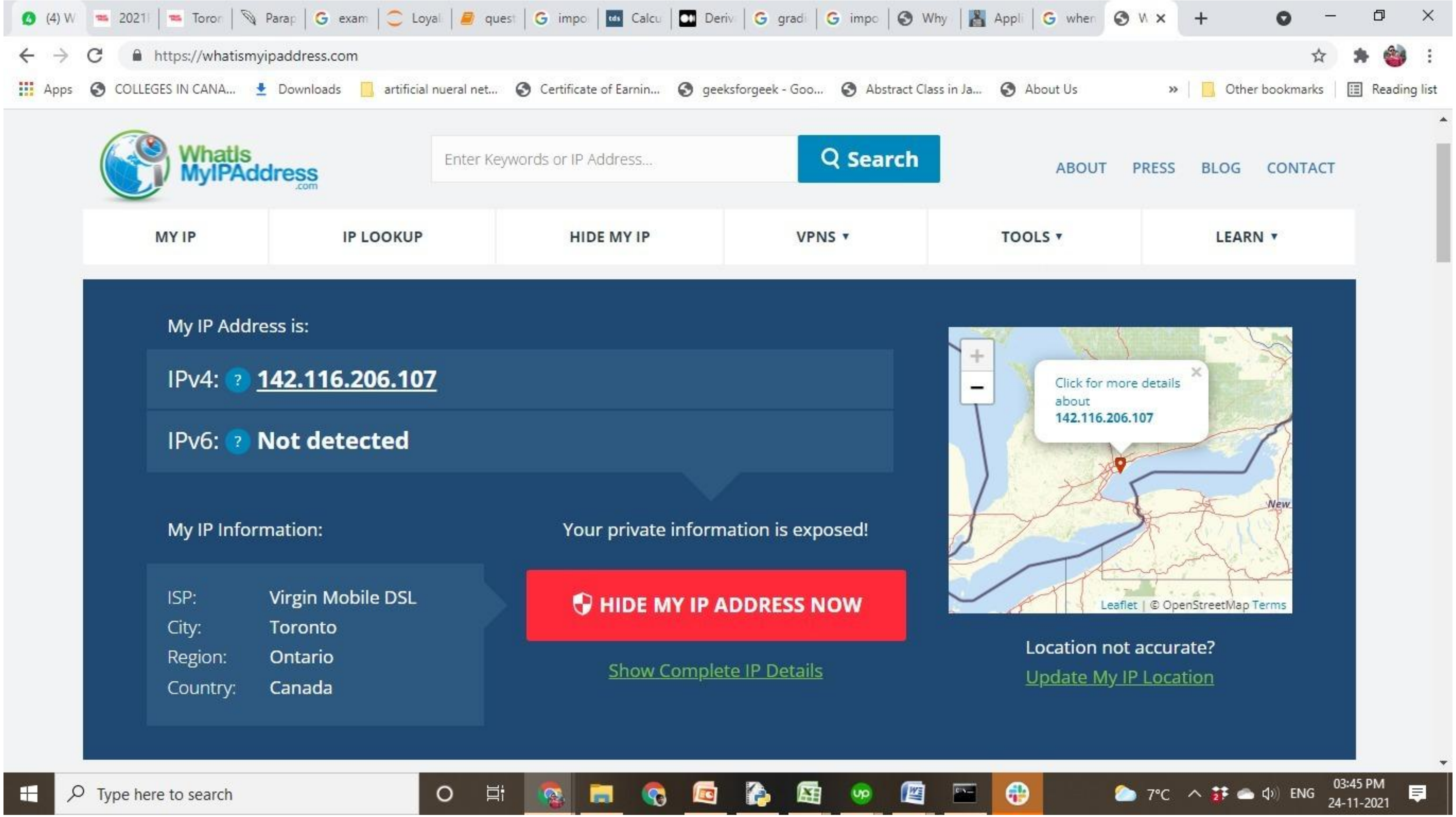
Personal Analysis – Data Cleaning and Modelling

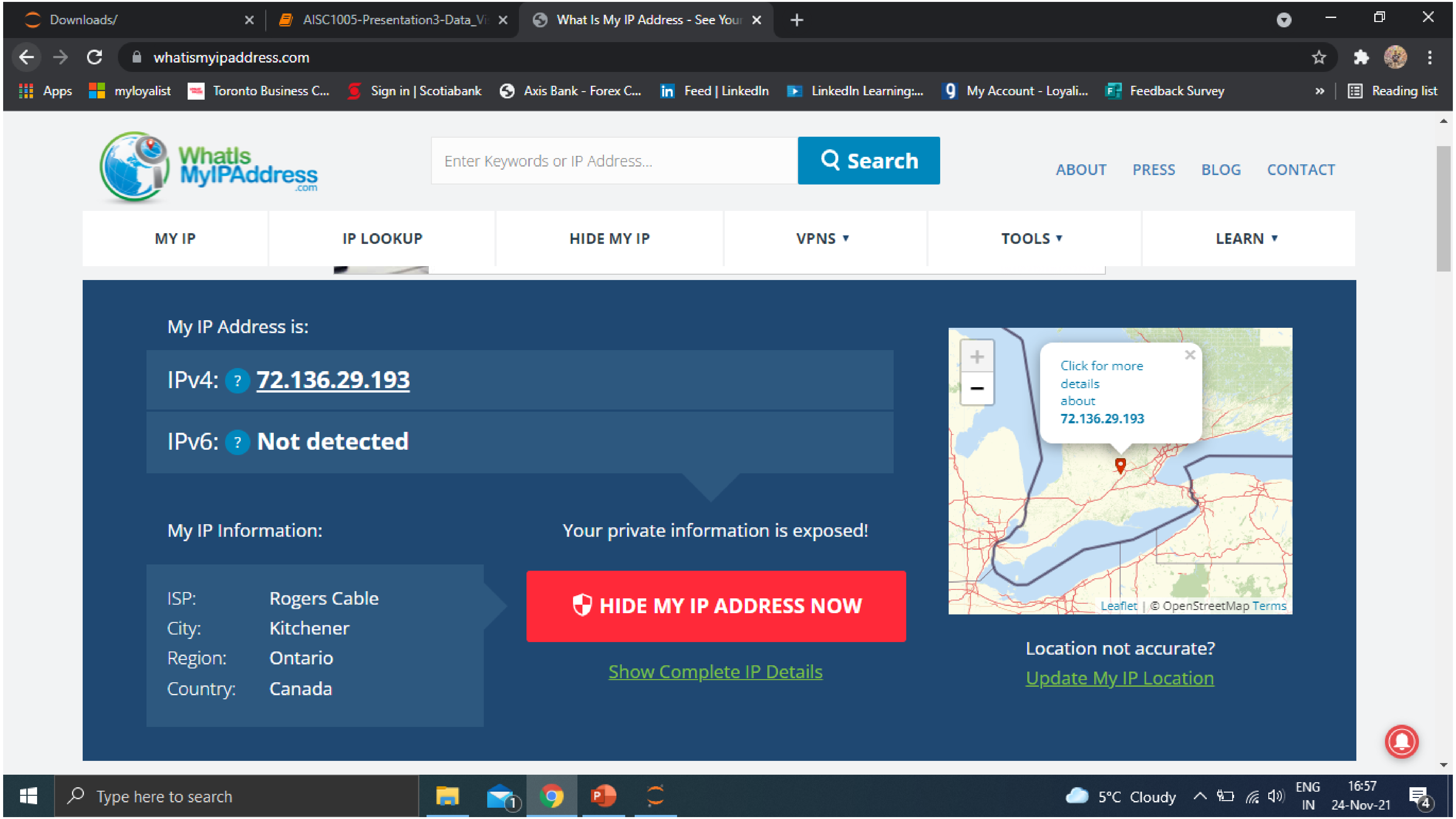
- After completing this project, I learn lot about data cleaning and modelling.
- We can improve this by using feature Selection, which help us in finding the most relevant feature to the house price.
- To improve accuracy, we can try some more regression techniques such as Support vector regression, Decision trees and Deep learning models.
- Further, we can deploy our project on any free server and make basic UI which takes all the parameters from the user and predicts the price of the house.

Personal Analysis – Data Visualization

- As we have created our own toy dataset, after data cleaning process, we analyzed the data and performed data visualization on it.
- We visualized the main information in form of pie charts, boxplot, scatterplot, histogram, etc.
- We found popular insights while performing data visualization on the dataset.
- We compared different features of houses accordingly to obtain accurate results.
- While doing this project, we understood that data visualization can make the audience understand the data in an effective way.
- In short, we cut the noise in our data and use only the useful patterns and values that can impact the business.


Saved to this PC
whatismyipaddress.com
WhatIs MyIPAddress .com
Enter Keywords or IP Address...
Search
ABOUT
PRESS
BLOG
CONTACT
MY IP
IP LOOKUP
HIDE MY IP
VPNS
TOOLS
LEARN
My IP Address is:
IPv4: ? 142.116.206.107
IPv6: ? Not detected
My IP Information:
Your private information is exposed!
ISP: Virgin Mobile DSL
City: Toronto
Region: Ontario
Country: Canada
HIDE MY IP ADDRESS NOW
Show Complete IP Details
Click for more details about 142.116.206.107
Leaflet | © OpenStreetMap Terms
Location not accurate?
Update My IP Location
Which is your biggest concern about using the Internet?
Type here to search
6°C
ENG
9:56 PM
2021-11-24


https://whatismyipaddress.com
Apps
COLLEGES IN CANA...
Downloads
artificial nueral net...
Certificate of Earnin...
geeksforgeek - Goo...
Abstract Class in Ja...
About Us
Other bookmarks
Reading list
WhatIs MyIPAddress .com
Enter Keywords or IP Address...
Search
ABOUT
PRESS
BLOG
CONTACT
MY IP
IP LOOKUP
HIDE MY IP
VPNS
TOOLS
LEARN
My IP Address is:
IPv4: 142.116.206.107
IPv6: Not detected
My IP Information:
ISP: Virgin Mobile DSL
City: Toronto
Region: Ontario
Country: Canada
Your private information is exposed!
HIDE MY IP ADDRESS NOW
Show Complete IP Details
Click for more details about 142.116.206.107
Leaflet | © OpenStreetMap Terms
Location not accurate?
Update My IP Location
Type here to search
7°C
ENG
03:45 PM
24-11-2021


Downloads/
AISC1005-Presentation3-Data_Vi
What Is My IP Address - See Your
whatismyipaddress.com
Apps
myloyalist
Toronto Business C...
Sign in | Scotiabank
Axis Bank - Forex C...
Feed | LinkedIn
LinkedIn Learning:...
My Account - Loyali...
Feedback Survey
Reading list
WhatIs MyIPAddress .com
Enter Keywords or IP Address...
Search
ABOUT
PRESS
BLOG
CONTACT
MY IP
IP LOOKUP
HIDE MY IP
VPNS
TOOLS
LEARN
My IP Address is:
IPv4: 72.136.29.193
IPv6: Not detected
My IP Information:
Your private information is exposed!
ISP: Rogers Cable
City: Kitchener
Region: Ontario
Country: Canada
HIDE MY IP ADDRESS NOW
Show Complete IP Details
Click for more details about 72.136.29.193
Leaflet | © OpenStreetMap Terms
Location not accurate?
Update My IP Location
Type here to search
5°C Cloudy
ENG IN
16:57
24-Nov-21

Thank You