**GTA Affordable Housing**

**Data Analysis**

**Kijiji.ca Rental**

1. **Introduction:**

The real estate price in GTA is a very ‘hot’ topic in any Medias, people are anxious to know the factors that decide the price and their price trend in next few years. Our project gathers real estate data from popular websites such as Craigslist, Twitter, Kijiji.ca and even GDELT data. Our goal is to research the data from different resources to summarize the complex factors that may influence the GTA real estate price.

1. **Data Source: Kijiji.ca rental**

Kijiji.ca is one of the largest Canadian online trading website which is focusing on local trading services. It has plenty real-time rental and real estate selling information and most are posted directly by individual owners.

1. PRO: real-time, direct price data from owner, local, categorized by region
2. Con: unformatted data, duplicated, unverified
3. **Data Gathering and ELT:**
4. I am using a Kijiji data scraper that downloaded from Github.

https://github.com/CRutkowski/Kijiji-Scraper

This scraper: **Kijiji-Scraper.py** is built on a popular Python library: Beautifulsoup 4. It parses and captures all AD url and some high level general AD information from the Kijij url you pass to it. It does not capture all detailed data from individual Kiji AD webpages.

1. **Kijiji-Scraper\_s.py** To capture the detailed data from individual Kijiji AD web pages, I modified the original scraper, rewrote the "ParseAd" function so it can scrape data from single AD web pages. It stores all data into a .txt file in dictionary format. It gathers detailed data, such as AD postdates, address and number of bathrooms.
2. **main.py** in this script **Kijiji-Scraper.py**, I called the scraper and pass over the parameters, to create txt files.

Including:

kijiji\_GTA.txt (any types of rental AD in GTA);

kijiji\_GTA\_basement.txt (basement rental AD in GTA);

kijiji\_GTA\_apt.txt (apartment)

kijiji\_GTA\_house.txt (house)

kijiji\_GTA\_one\_bed.txt (one bed room)

kijiji\_GTA\_two\_bed.txt (two bed room)

1. **Main\_scr\_all\_pages.py**. This script calls **Kijiji-Scraper\_s.py** to scrape all detailed AD data from the AD urls we gathered from main.py (step 3). Due to Kijiji's span protection, Kijiji block my IP address when my requests reach 2000. So I break down the total urls into smaller group, each group has about 2000 urls. It store data in a few txt files.
2. **Convert\_CSV2.py** Read all the txt files that created from Main\_scr\_all\_pages.py, append them into one dataframe file, drop duplicated on AD\_ID, then store into a csv file.
3. **Insert\_Table.py** Insert all data into a Postgres table, hosted on Google Cloud. It connects to **Google cloud** on Proxy protocol, use psycopg2 connects to the cloud database.
4. **Download and clean data:**

**Project.ipynb** connect to **Google Cloud**, Postgres database, download all the rental data from gtarent\_s table.

1. Sort data on postdate:

df2 = df.sort\_values('postdate', ascending=True)

1. Clean “Price”, remove ‘$’ and null

df\_st1.price= [x.replace('$','') **for** x **in** df\_st1.price]

df\_st1.price= [y.replace(',', '') **for** y **in** df\_st1.price]

1. Clean “adbedrooms”, remove space, ‘Studio’ and ‘Bachelor’

df\_t['adbedrooms']=[z.replace('+','') **for** z **in** df\_t['adbedrooms']]

df\_t['adbedrooms']=df\_t['adbedrooms'].replace('', 0)

df\_t['adbedrooms']=df\_t['adbedrooms'].replace('Studio', 0)

df\_t['adbedrooms']=df\_t['adbedrooms'].replace('Bachelor', 0)

1. Clean “adbathrooms”, remove spaces:

df\_t1['adbathrooms']=df\_t1['adbathrooms'].replace('',0)

1. Clean “geolan”:

df\_t3['geolan']=[z1.replace('+','') **for** z1 **in** df\_t3['geolan']]

df\_t3['geolan']=[z2.replace(',','') **for** z2 **in** df\_t3['geolan']]

df\_t3['geolan']=df\_t3['geolan'].str.strip().replace('',0)

1. Clean “geolng”:

df\_t4['geolng']=[z1.replace('+','') **for** z1 **in** df\_t4['geolng']]

df\_t4['geolng']=[z2.replace(',','') **for** z2 **in** df\_t4['geolng']]

df\_t4['geolng']=df\_t4['geolng'].str.strip().replace('',0)

1. Convert “price”, “adbedrooms”,” adbathrooms”, “geolan” and “geolng” into FLOAT type
2. Data Analysis:
3. Geo Data Analysis:
4. Conclusion: