BUAN 6385
ROBOTIC PROCESS AUTOMATION

AUTOMATING REAL ESTATE WEB SCRAPING USING UIPATH

NAME: KAUSHIK MURALI
NETID: KXM230003

PROBLEM STATEMENT

Searching for homes, properties for rent or sale is a tedious task and, in this project, we are trying to automate the process of Real Estate Web Scraping.

ABSTRACT

In this project, we make it easy to collect data from a variety of real estate websites. Web scraping is the process of acquiring vital information such as property data, pricing, and locations to compare or use elsewhere. We may use UiPath, a popular task automation tool, to construct bots that search websites, extract information, and store it neatly in spreadsheets or databases. What is the goal? To improve the speed with which we receive real estate information to make more timely decisions.

OBJECTIVE

Our primary purpose is to make it easier to acquire key real estate information such as property listings, pricing, and descriptions from a variety of websites. We are automating the search with UiPath RPA to construct a large property database. This database will allow us to compare data, evaluate it, and potentially combine it with other systems to make better real estate decisions. In addition, we want to distribute this information via email and WhatsApp to keep everyone up to date and operate it seamlessly.

INTRODUCTION

Real estate web scraping with UiPath means using special tools and programs to gather important info from lots of real estate websites. UiPath Studio, which is like a toolbox for this kind of work, gives us everything we need to do it well. First, we pick the website we want to scrape and set up our project. Then, we make sure we have all the right tools installed and use an option called "Open Browser" to get onto the real estate site.

Once we are on the real estate website, UiPath assists us in finding the information we need, such as property details and pricing. If there are a lot of pages, we find out how to go through each one and capture all the information.

After we have gathered all the necessary information, UiPath allows us to save it in a variety of formats, including Excel and CSV files. To ensure that everything runs properly, we prepare for any difficulties that may arise during scraping, such as mistakes or website troubles. Once we have tried everything and confirmed that it works properly, we can arrange it to run automatically whenever we want. But it is super important to play by the rules of the website we are scraping. We must be careful not to take too much data or do anything that is not allowed. In the end, UiPath makes real estate web scraping much easier, helping us get the important info we need quickly and without any hassle.

PROJECT JUSTIFICATION

House hunting is a drag. Taking time away from work, family, and other obligations to search for a house in the paper or online and drive around scheduling open houses can be a real challenge.

This project will make finding a house easier by:

1. *Automated Data Scraping:* Instead of searching websites and copy pasting information from them, the project will scrape data from these sites automatically, including property description, price and location. This will save you a lot of time and effort.
2. *Better efficiency:* All the properties for sale would be there, all in one place, at a click of your mouse, and all the information you would ever need about each one. The listings would be gathered from different sources, and you would be able to see briefly which ones meet your requirements.
3. *Error Reduction:* Data entered manually can sometimes have minor typos or otherwise be incorrect. This project takes data directly from online listings so your results are as accurate as possible.

This project aims to make it faster, easier, and more efficient for people to find a house.

SELECTION CRITERIA

The RPA is our secret weapon that enables us in the real estate data.

Now think of the situation where you are on the internet searching through real estate websites, copying, and pasting the data repeatedly. It is a very boring job, and it is very difficult to avoid mistakes. This is the power of RPA (Robotic Process Automation) that helps us win the battle. Consider it as an unflagging assistant who can be relied upon to extract data in large quantities and high complexity without any human intervention. Without RPA it would be hard to complete the task on time and with accuracy.

Modernization: Much laboring has been replaced by machines.

In the old days, looking for houses was a very time-consuming process. It is not fast, more error prone, and even cannot manage the growing number of websites and data points. The purpose of this project is to make the process more efficient by using automation. It can be thought of as a smart upgrade that replaces monotonous work with a more accurate and faster system.

Stability: A Sound Basis for Consistent Results.

We are creating a project on a strong foundation, like a solid house base. This platform, together with a separate tool for Microsoft Edge, can be applied to the most complex sites with enormous amounts of data. It is a reliable choice to work with as it is certain to provide dependable results even on difficult sites, without any glitches.

Adapting to Change: Being in the same step with the market.

The real estate market is a dynamic environment, and so this project should adapt accordingly. We can always do better and enhance the automated search method to always be at the cutting edge. This adaptability enables the system to respond to the variations in web pages or data formats to keep your data up-to-date.

Quality Matters: Providing leading data extraction.

We believe that the data extraction system should be built as good as a house, and therefore we are committed to quality. This RPA solution is based on the best practices, which guarantee a continuous, accurate and errorless process. It implies that you can be certain about the data you receive and use it to make well-informed decisions.

Data Quality: Reliable Data Results in More Informed Decisions

Manual data entry is a pot of errors. RPA will automate the process and help us to collect data that has higher accuracy and consistency. This results in better quality data for analysis, which in turn, will lead to more informed decisions when trying to find the perfect home.

Tools for Efficient Extraction with a Big Impact

Imagine the tools we will use – UI Path Studio and the Microsoft Edge plugin – as our toolboxes for creating this automation system. They offer you a wide range of tools for handling any kind of situations that you may come across when scraping data online from the websites, such as dealing with dynamic content, logging in to the website, and data validation. This makes the procedure to be smooth and effective for the extraction process from any website.

STAKE HOLDERS

Through the examination of the web of concerned parties who are affected by Real Estate Data Scraping Automation, we can see that the technology comes with both positive and negative implications.

Web scraping is the technology that has a chance to disrupt the real estate industry, and to change it in the favor of the bulk of the participants. Here is a breakdown of how each group is interested in or impacted by this practice:

1. Real Estate Agents & Brokers:

- Benefits: Go through the extended selection of listings, track the real time market changes, find the potential clients, and give more accurate recommendations to the buyers and sellers.
- Challenges: They may also be displaced by tools that automate tasks such as lead generation. They may have to redesign their services to take advantage of data insights that provide more value to their clients.

2. Real Estate Investors:

- Benefits: Reduce time and effort for market research, find hidden opportunities, increase investment returns, make data-driven transaction, sales and portfolio management decisions.
- Challenges: Need skills to decode and use scraped data correctly, ponder on issues related to data usage, for example copyright and privacy.

3. Real Estate Developers:

- Benefits: Determine locations with the highest demand, evaluate if the new projects are profitable economically, and decide on location, type of property, and prices in accordance with the current market conditions.
- Challenges: The accuracy and completeness of data are fundamental, which means that all the data should be filtered and analyzed thoroughly to avoid the situation when decisions are taken based on incorrect data.

4. Property Managers:

- *Benefits:* Learn about rental market trends, apply pricing strategies wisely, discover vacancies quickly, and increase your occupancy rates with the use of data-driven decision making.
- *Challenges:* The main feature of the information accuracy is the right rent evaluation and the empty property managing. Fair housing is one of the major concerns of the community.

5. Financial Institutions:

- *Benefits:* The automated valuation models will do more accurate risk assessment for mortgages and loans, cut down on underwriting processes, and give out competitive rates based on data-driven property valuations.
- *Challenges:* Data protection and keeping to privacy laws are a must. The awareness of the deficiency of the scraped data and the requirement of the human intelligence for the financial decision-making.

6. Government Agencies:

- *Benefits:* Develop the informed urban planning strategies, make the decisions on zoning laws based on data, and the property taxes of real estate exactly evaluate using the comprehensive real estate data.
- *Challenges:* Data correctness and uniformity of different data sources are the problems. The issue of privacy while using public data collection should be taken seriously, and it is vital to address any privacy issues that may arise.

7. Data Analysts & Researchers:

- *Benefits:* The big data access for market research, trend analysis, and developing predictive models which can then be used to make informed decisions regarding the properties.
- *Challenges:* Data cleaning, filtering and validation are the critical steps to the reliable analysis. Knowing the constraints of scraped data and the need to integrate it with other sources for a bigger picture.

8. Technology Providers:

- *Benefits:* Incorporate web scraping as part of their software solutions for real estate, which will give their clients more advanced data-driven characteristics and functionality.
- *Challenges:* Following all the legal and ethical guidelines for web scraping, and making sure that their users' data is safe within their platforms.

9. Legal & Compliance Teams:

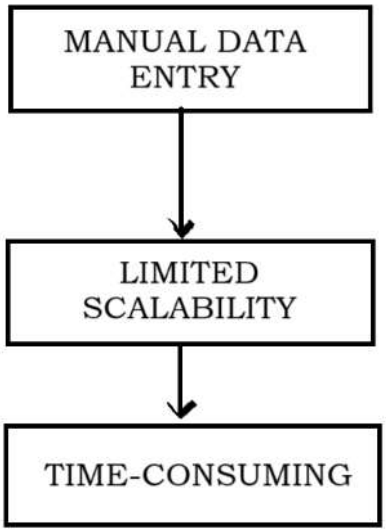
- *Role:* Inform stakeholders on the regulations of data privacy, website terms of service and intellectual property laws so that they can avoid legal issues linked to data scraping.
- *Challenges:* Web scraping legal ground is a dynamic field, and the compliance with the regional and international data privacy standards should be considered as a priority.

10. End Users (Buyers, Renters, Sellers):

- *Benefits:* Gain the possibility of viewing a wider variety of listings, compare prices, look for the property that meets your exact requirements and make wise decisions about buying, renting, and selling the real estate.
- *Challenges:* Large data sets, incorrect or misleading listings, and the necessity to master filtering and evaluating skills of scrapped data are the main challenges.

AS-IS WORKFLOW

The AS-IS workflow diagram shows a manual process to capture property data for a real estate property and enter it into a system, typically a database or a property management software. Here is a breakdown of the potential steps involved:



1. Data Gathering:

Source Identification: Agents or data entry people go through different sources to collect data about properties. This could include:

- Physical visitations to the properties for collecting the details like square footage, number of bedrooms and bathrooms, amenities, and overall health condition.
- Phone calls or emails with property owners or sellers to ask about the asking price, details about the listing and availability.
- Real estate paper documents or marketing materials with property details that must be entered manually into the system.

2. Manual Data Entry:

- *Data Input Platform:*

  The collected information is subsequently typed into a separate system by hands. This could be a standalone property management application, a spreadsheet, or a web-based form.

- *Data Fields:*

  The fields that one needs to fill may not be the same in all systems and it also depends on the type of property. Examples of common fields include:

  - *Property Details:* The information consists of address, property type (single-family home, apartment, condo), square footage, number of bedrooms and bathrooms and year built.
  - *Listing Details:* The listing should include the asking price, listing type (for sale, for rent), availability date, description of the property's features and amenities.
  - *Agent Information:* Agent or agency contact information on the listing, which is responsible for that property.
  - *Additional Information:* The notes, photos (can be uploaded separately), virtual tour links and any other information about the property that might be of interest.

3. Data Review:

The process may differ according to the workflow. Therefore, the first step may involve a basic review process to identify any obvious errors or inconsistencies in the entered data. Nevertheless, it might be conducted by the manual process that is not so detailed.

4. Data Output:

The data is uploaded and stored (if approved) into the chosen system after the users enters and (possibly) reviews the data. This enables the handling of electronic property listings to the basic level.

5. Limitations of AS-IS Workflow:

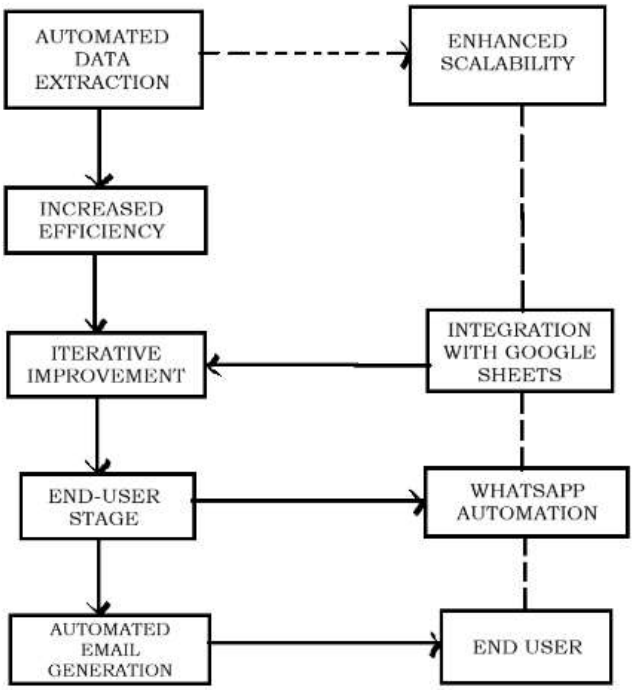
- *Time-Consuming:* Manual data entry is a slow and tedious process that really slows down the process if there are many properties.
- *Error-Prone:* Typing fault, wrong format in data, and missing details in the entered data are some of the examples of errors that occur during manual entry.
- *Limited Scalability:* When the number of properties is growing, the data entry workload for manual operations becomes unbearable.
- *Lack of Automation:* The automation is simply for jobs like image uploads, data validation, description generating.

This AS-IS workflow focuses on data entry and is filled with human errors and inefficiency. Technology will help in the effective workflow optimization and will make property management processes much better for real estate professionals.

TO-BE PROPOSED WORKFLOW

The workflow is improved with multi-channel communication.

Automation of data delivery and communication with the end-users (potential property buyers, renters, or investors) are the main advantages of this workflow that has been developed during data processing stages.



1. Data Preparation:

The first steps will most probably involve data retrieval from different sources (web scraping, databases, spreadsheets). This data goes through the process of integration where information from different sources is consolidated into one format. Iterative improvements may include data cleansing, the elimination of errors, and checking the consistency of data.

2. End-User Stage:

This is the time when the processed real estate data reach the target audience.

3. Multi-Channel Communication:

- *WhatsApp Automation:* The system integrates with WhatsApp to send out timely updates, notifications, or messages to the end users directly on WhatsApp platform. This is the reason why the data delivery becomes more conversational and interactive in nature.

- *Automated Email Generation:* The detailed reports and comprehensive information are sent to end users via e-mail. Emails can be used to send more elaborate data visualizations, property details, or disclaimers instead of texts.

Benefits of Multi-Channel Approach:

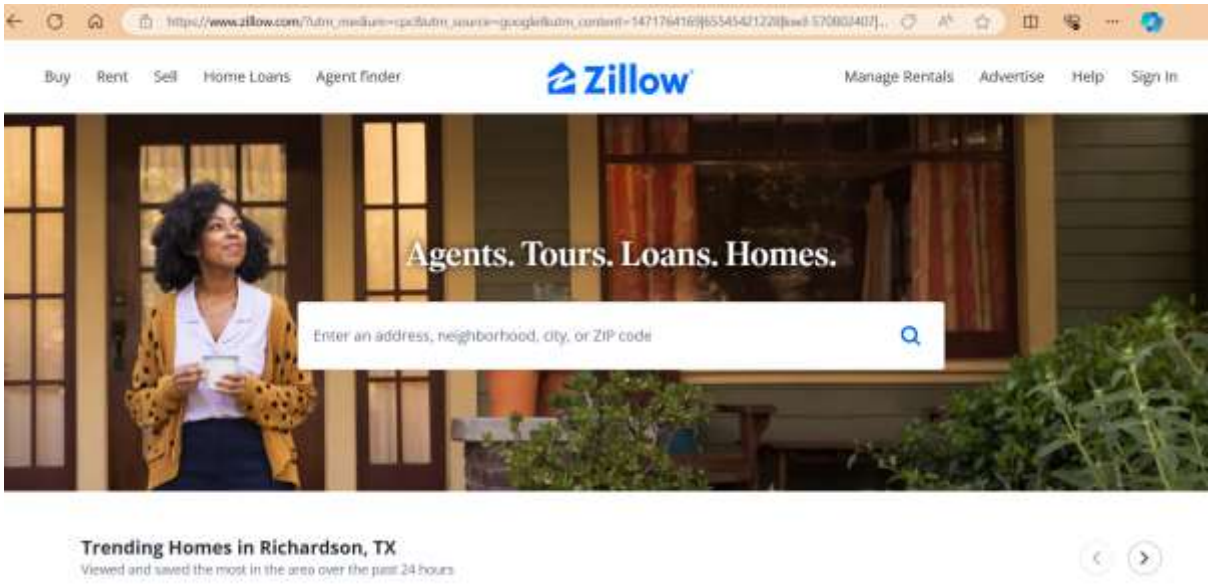
1. *Increased Engagement:* Through both WhatsApp and email, information is being made accessible to a larger number of users with different communication preferences. Some users may find it convenient to get some updates on WhatsApp while others may prefer the detailed reports that can be emailed.
2. *Improved User Experience:* The user is served with the right information in time through their preferred channels; this makes the process of engaging with real estate data enjoyable.
3. *Targeted Communication:* The system's capability of providing data and notifications can be customized to a certain point, depending on whether the user is interested in a specific property or not.

As a result, this TO-BE workflow is intended to simplify data delivery and exploit automation to enhance communication and users' engagement in the real estate market. The real estate data will be available through WhatsApp and email and this will enable the users to obtain the information they need instantly and comfortably.

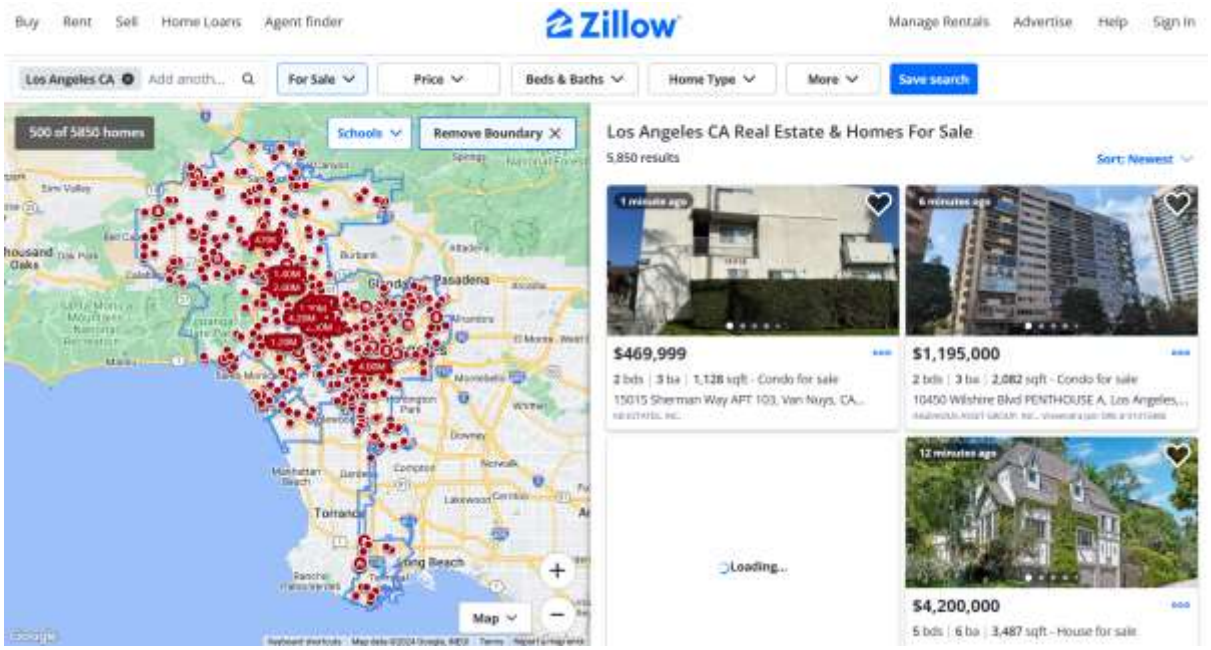
VIEW OF REAL ESTATE DATA

In this project, I have used Zillow website as my platform to search for real estate properties.

When you open the Zillow website, it looks like this:



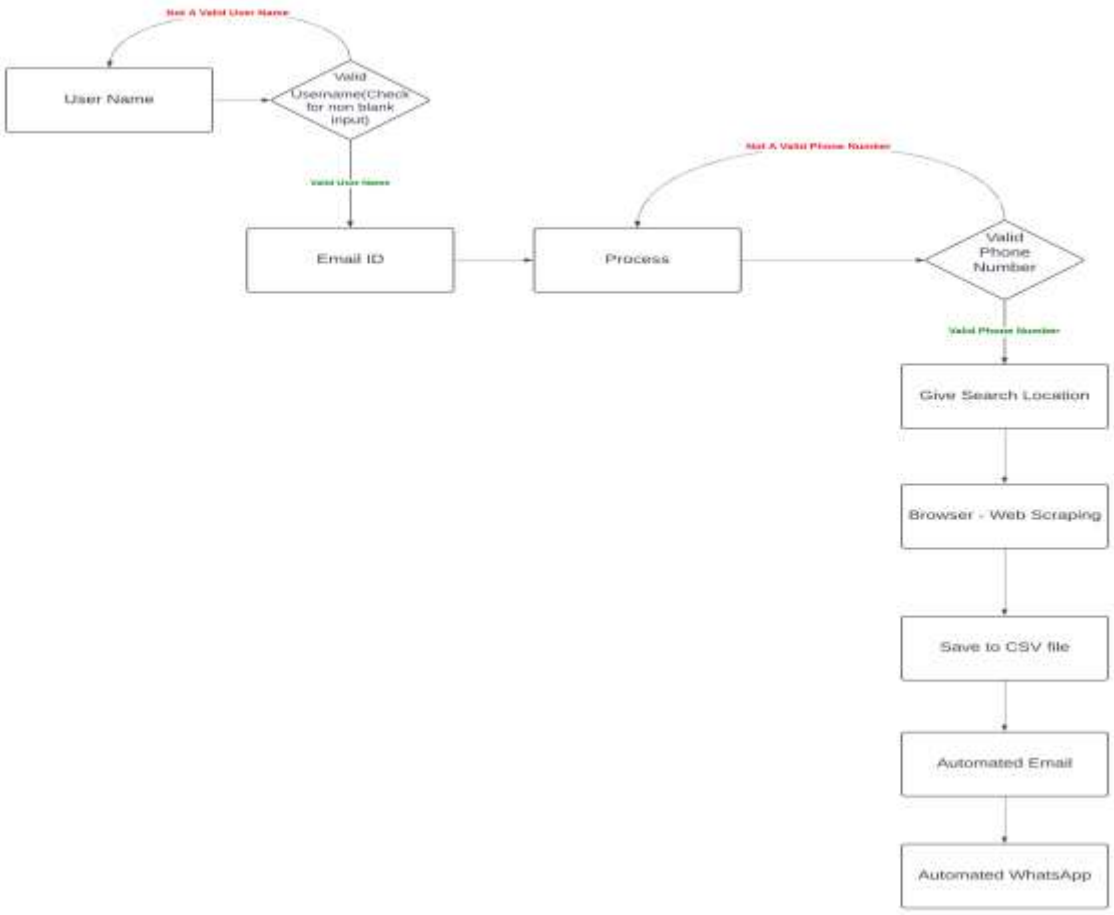
What we usually do is go to the website, enter an address and search for properties. We then choose the newest listings from the dropdown. The search engine returns results like this:



So, instead of scrolling down and looking into every property detail, with the RPA project, I extract the main information, like the pricing, the address, image link, number of bedrooms, number of bathrooms and built area.

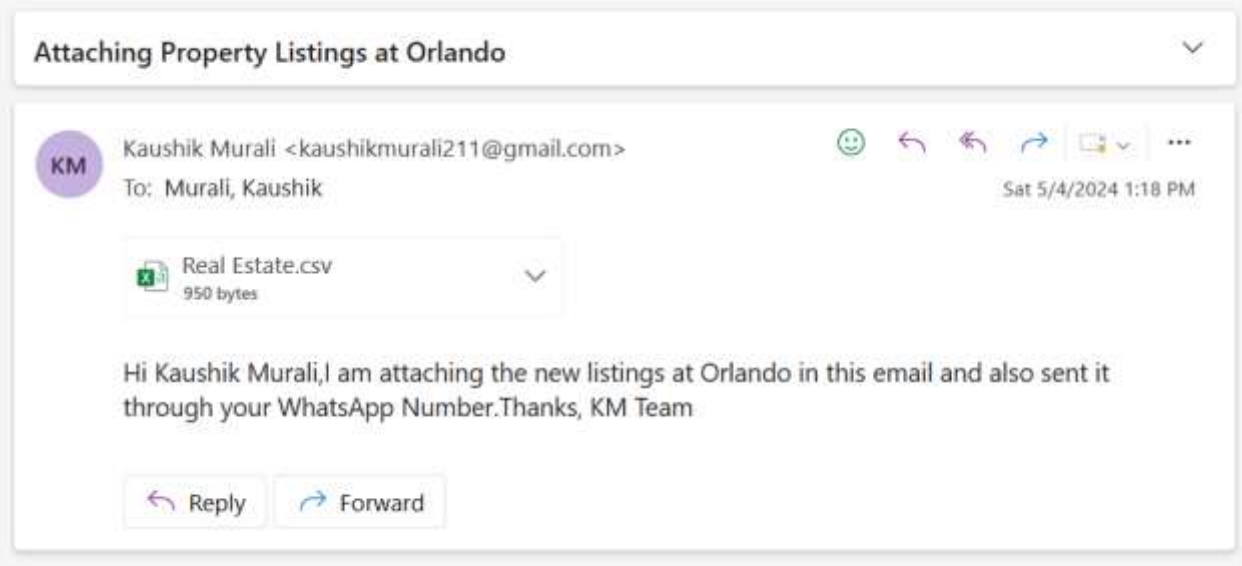
CURRENT MODEL WORKFLOW:

In the current RPA model, the model starts with asking input about User Name, checks whether the given input is null or not. If the input is not null, then it goes to obtaining the email ID of the user. Followed by that, the RPA gets input of the phone number with the country code and if it is a valid phone number the user gets to give the input for the place of interest where he/she is looking for a property. Once the automation gets a valid input, it opens the Microsoft edge browser and accesses the Zillow website and web-scraping starts here by utilizing the extract table function.

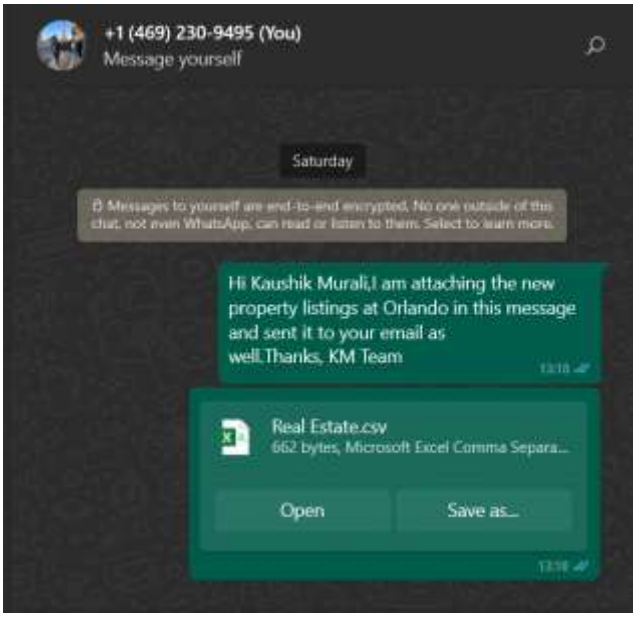


Once the data – the price, the area of property, number of bedrooms, number of bathrooms, address, and the link for the property images, are extracted, they get stored into a CSV file on the local system. Now, this CSV file gets attached and is sent through email and WhatsApp. So, overall, we have got 4 automations. Microsoft Edge application opening, Excel Automation, Email Automation and WhatsApp Automation.

Sample output for Email automation is given below:



Sample output for WhatsApp automation is given below:



This is the output for message sent from my account to myself as I typed in my phone number as the input.

TESTING PLAN

An exhaustive testing plan cannot be understated as it is essential for achieving a running RPA model without any glitches. The functional test should start with user input validation, error handling of null values, invalid characters, as well as wrong phone number entry formats in usernames and phone numbers. The webpage structure detection should be tested for all Zillow webpages to make sure it is adaptive to layout alterations. Moreover, checking the accuracy of scraping all needed data fields (price, area, bedrooms, bathrooms, address, and image link) is essential as well. Data storage test should be performed with the purpose of verifying that all extracted data is saved correctly into the CSV file and by handling possible errors that occur while creating or saving the file due to lack of disk space. Delivery testing for notifications should involve various email providers to ensure that the emails are compatible, are delivered successfully with the CSV file attachments, and use the WhatsApp message delivery option. It is also necessary to test the notification content to ensure that the data displays correctly in both emails and WhatsApp messages.

The integration testing should be concentrated on the verification of the seamless coexistence of different RPA parts. For instance, we have user input validation, scraping initiation, successful scraping, storage of the CSV file, data retrieval for email/WhatsApp notification creation, and successful attachment of the CSV file to emails and its inclusion in WhatsApp messages. In the end, the general workflow should be tested through a trial run from start to finish checking the entire process from user input to notification delivery.

In addition to the testing on a functional level, the testing on a non-functional level should be accounted for also. Time performance of the entire process with various data volumes should

be tested to determine the relationship between the number of properties scraped and processing speed. Address the effect of network speed on scraping efficiency as well. Privacy checks must be conducted if the model processes user data. Provide a secure storage and transmission of data, and check for the possibility of web scraping and data transmission.

To test the error handling and recovery abilities, it is essential to check how the model behaves under unpredictable conditions such as website unavailability, network problems, or incorrect data formats. Ideally, the program should resume execution or provide useful error messages after recovering from errors. In addition, the last but not the least thing to do is to do regression testing after any modification or update of the model. This means that the features that already work would not be affected and the old problems caused by the new issues would not reappear.

Through this inclusive testing plan, you will be able to predict the issues, revise your RPA model, and confirm that it is fully operational and successful in the data scraping and notification processes of real estate automation.

CURRENT MODEL DRAWBACKS

1. **User Name input and Valid place input test is only done for blank input:**

   This implies that the system is not checking for the correct input from the user. For instance, if you enter an "X" as your username, the system should recognize it as an invalid username and the system should prompt you to enter a valid username. Nevertheless, this scenario is not valid as "X" is an invalid username and the system lets you move to the next step without any validation. This may cause some security issues or wrong information being processed.

2. **WhatsApp automation works only for US numbers:**

   This means that the phone numbers that can be handled by the WhatsApp automation of the system are limited to those in the United States. If you attempt to use a phone number from another country, the automation may not function properly or may not work at all. This limitation might reduce the applicability of the system for users outside the United States and will need further programming to support international phone numbers.

3. **This model works with only Zillow website. If we want to check for other websites, the user must change the website in the model:**

   The system is developed for the purpose of collecting data from the Zillow website exclusively. For instance, if you want to use Zillow or Trulia, you might be required to tweak the system to work with these websites. This could require the transformation of the way the system works through the website, data extraction and dealing with the website-specific elements. The users will require programming skills or support from technical personnel to modify the system in an efficient manner.

In sum, the mentioned points reflect the areas where the system needs to be adjusted or developed to increase the functionality, usage, and flexibility.

PROJECT DELIVERABLES

The project is targeting the creation of an automated web scraping tool that extracts real-time data from several real estate websites. This application runs autonomously, which simplifies the process and eliminates the need for human interference.

By employing high level web scraping methods, the platform collects detailed property datasets including listings, prices, locations, descriptions, and images. These data sets are designed in a way that will help in the analysis, comparison, and integration with other applications or systems if needed. Consumers can employ the data for a thorough analysis, identify the market tendencies and take smart business decisions.

Moreover, the tool provides smooth integration functions, allowing the users to import the data into different applications or systems to ensure better functionality.

In the end, the project is intended to give users access to reliable and up-to-date real estate data, which will then be used to increase the efficiency and decision-making process of the real estate industry.

TOOLS USED

In the process of web scraping which is facilitated using UI Path Studio and the Microsoft Edge plugin, these tools are intertwined and are designed to automate the process of taking data from real estate websites. The back-end of the system has been combined with CSV sheet that provides us with a convenient platform for storing and managing the collected real estate information. The collaboration of UI Path studio, Microsoft Edge plugin and CSV

sheet, makes it possible to carry out this procedure in a smooth manner for data extraction, storage, and analysis tasks. By this integrated approach, the solution increases the efficiency and scalability, thus smooth operations can be facilitated all the way through the data life cycle.

FUTURE IMPROVEMENTS

1. *User Experience:* Provide a choice to users for notification channels (email/WhatsApp) and search filters for data collection that is specific to the user. Give updates on progress during the phase of implementation.

2. *Data Processing:* Include external providers in the data enrichment, create visual reports, and implement the anomaly detection for the unusual data points.

3. *Scalability & Reliability:* In this regard, take advantage of cloud computing for better data management and storage. Develop efficient error handling and recovery procedures. Let the program run by itself with scheduling for automatic data scraping and delivery.

4. *Security & Compliance:* Data security should be given priority by application encryption and access controls. Stick to the latest data privacy regulations.

5. *Integration:* Find out how to combine our software with the existing real estate software for seamless workflow.

REFERENCES

1. https://forum.uipath.com/t/how-to-scrape-in-nested-data/356583
2. https://forum.uipath.com/t/how-to-extract-data-from-website-using-dynamic-data-scraping/238655
3. https://docs.uipath.com/activities/other/latest/integration-service/uipath-twilio-messaging-send-whatsapp-message
4. https://ijrpr.com/uploads/V4ISSUE6/IJRPR14525.pdf