Data Crawling and Collecting

PROJECT SUBMITTED TO ASIAN SCHOOL OF MEDIA STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF

M.Sc. in Data Science

By

Ankit Pawar

(University Enroll. No: MS-22543)

Under the Supervision of

Prof. Mr Nitish Patil



2024

DECLARATION

I, Ankit Pawar, S/O Pan singh Pawar, declare that my project entitled "Data Crawling and Collecting", submitted at School of Data Science, Asian School of Media Studies, Film City, Noida, for the award of M.Sc. in Data Science, Noida University and Post Graduate Data Science, ASMS, is an original work and no similar work has been done in India anywhere else to the best of my knowledge and belief.

This project has not been previously submitted for any other degree of this or any other University/Institute.



Signature

Ankit Pawar 88106262302 Ankitpawar483@gmail.com M.Sc. Data Science School of Data Science

ACKNOWLEDGEMENT

The completion of the project titled "Data Crawling and Collecting", gives me an opportunity to convey my gratitude to all those who helped to complete this project successfully. I express special thanks:

- To *Prof. Sandeep Marwah*, President, Asian School of Media Studies, who has been a source of perpetual inspiration throughout this project.
- To *Mr. Ashish Garg*, Director for School of Animation/Data Science for your valuable guidance, support, consistent encouragement, advice and timely suggestions.
- To *Mr.Nitish Patil*, Assistant Professor of School of Animation/Data Sciecne, for your encouragement and support. I deeply value your guidance.
- To my *faculties & friends* for their insightful comments on early drafts and for being my worst critic. You are all the light that shows me the way.

 To all the people who have directly or indirectly contributed to the writing

of this thesis, but their names have not been mentioned here.

Signature

Ankit Pawar 8810626302 Ankitpawar483@gmail.com M.Sc. Data Science School of Data Science

Table of Contents

Contents	Page No.
About the Company	5
About Job Role	6
Week-1	8
Week-2	9
Week-3	10
Week-4	11
Week-5	12
Week-6	13
Week-7	14
Week-8	16
Week-9	17
Week-10	19
Week-11	20
Week -12	21
Conclusion (Learning Outcome)	23
Ribliography	24

About the Company

Blackopal Group

Embark on an exciting real estate journey with us! Our mission is simple: Customer Empowerment. We achieve this through smart tools, perfect match-making, and customized credit solutions. Bid farewell to the buying confusion.

To bring excellence to our mission, we attract, nurture, and retain the best talent. Our organization thrives on creativity and ideas, guided solely by meritocracy. From a mere idea on the drawing board, we have grown into a team of over 50+ ambitious professionals who share the dream of helping customers find their dream homes every single day.

Our ultimate goal? To revolutionize the real estate experience worldwide. We're not satisfied with the status quo. Instead, we aim to redefine how we make buying decisions. Together, let's usher in a new era, transforming the way we experience real estate and unlocking endless possibilities.

About Job Role Designation - Market Research Analyst Intern

- Data Extraction Specialist skilled in Selenium foweb scraping, adept at meticulous data cleaning using Pandas and Numpy.
- Performing visualizations with Matplotlib to communicate actionable insights effectively.
- Proven track record in extracting real estate data from diverse RERA portals and enhancing data accuracy through advanced cleaning techniques.
- Developing an insights section in the Application for showcasing price trends and construction cost graphs of the real estate projects.
- Enabling users to track and analyze the fluctuation over time.

Week-1 Report

Date 5 Feb	From	To 9 Feb
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Descriptions:-

Data Collection:

- Gather price trend data from your clients. This could include historical prices, current prices, and possibly projected future prices.
- Organize this data in a structured format. Each row could represent a specific time period (e.g., month) and each column could represent a different product or market segment.

Record Keeping in Excel:

- Set up an Excel sheet with appropriate column headers to store the collected data. Include columns for date, product name, price, and any other relevant information.
- Enter the collected price trend data into the Excel sheet.

Data Analysis:

- Once you have a sufficient amount of data collected, you can start analyzing it.
- Utilize Excel functions and tools to calculate metrics such as average prices, price changes over time, and price volatility.
- Create charts and graphs to visualize the trends in the data. This can help in identifying patterns and outliers.

Week-2 Report

Date 12 Feb	From	To 17 Feb
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Descriptions:-

Comparison Model:

- Develop a comparison model to analyze the price trends in the micro market.
- This model could involve comparing the price trends of different products or market segments within the micro market.
- Consider factors such as demand, supply, competition, and external economic factors that may influence prices.
- Utilize statistical techniques or machine learning algorithms if necessary to uncover insights from the data.

Interpretation and Reporting:

- Interpret the findings from your analysis and comparison model.
- Identify any significant trends, anomalies, or patterns observed in the data.
- Prepare a report summarizing the analysis and providing insights into the micro market's price dynamics.
- Present the findings to relevant stakeholders, such as clients or management, in a clear and understandable manner.

Regular Updates:

- Continuously update the Excel sheet with new price trend data as it becomes available.
- Regularly rerun the comparison model to keep track of changes in the micro market over time.

Week-3 Report

Date 19 Feb	From	To 24 Feb
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Descriptions:-

Collecting property data from a government site using Selenium and then cleaning the data can be achieved through a series of steps. Here's a general outline of how you can approach this task:

1. Setting Up Selenium:

- Install Selenium library using pip if you haven't already: Pip Install selenium.
- Download the appropriate web driver (e.g., ChromeDriver for Google Chrome) and ensure it's in your system PATH.

2. Automating Data Collection:

- Use Selenium to automate the process of navigating to the government site, entering search criteria (if any), and scraping the property data.
- Identify the elements on the website (such as input fields, buttons, and table rows) that contain the data you need and use Selenium to interact with them.

3. Scraping Property Data:

- Write code to scrape the relevant property data from the website. This could include information such as property address, size, price, amenities, etc.
- Store the scraped data in a suitable data structure, such as a list of dictionaries or a pandas DataFrame.

Week-4 Report

Date 26 Feb	From	To 2 Feb
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Descriptions:-

Storing the Cleaned Data:

• After cleaning the data, you can store it in a structured format for further analysis. Common options include saving it to a CSV file, Excel spreadsheet, or a database.

Iterating and Error Handling:

- Since you're collecting a large amount of data (over 800 records), it's essential to handle error gracefully and ensure the automation script can resume from where it left off in case of interruptions.
- Implement error handling mechanisms and logging to track any issues encountered during the data collection and cleaning process.

Testing and Validation:

- Before finalizing the data, it's a good practice to perform testing and validation to ensure the accuracy and integrity of the collected and cleaned data.
- Manually review a sample of the data to check for any discrepancies or errors that may have been overlooked during the cleaning process.

By following these steps, you can use Selenium to collect property data from the government site, clean it to prepare it for analysis, and then proceed with further processing or analysis as needed.

Week-5 Report

Date 4 Mar	From	To 9 Mar
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Descriptions:-

One such area is real estate, where property data can provide valuable insights for investors, developers, and policymakers. This explanation delves into the process of collecting and cleaning property data from government websites for Noida and Ghaziabad using the Selenium library.

The first step involves setting up the Selenium library, a powerful tool for automating web browser interactions. With Selenium, we can programmatically navigate government websites, enter search criteria, and extract property data. It's essential to ensure that the appropriate web driver, such as ChromeDriver for Google Chrome, is downloaded and configured correctly.

Once Selenium is set up, we proceed to write a script that automates the data collection process. By identifying HTML elements containing the desired data—such as property addresses, sizes, and prices—we can extract this information efficiently. Since we aim to collect over 800 data points, the script must iterate through multiple pages or search results to gather comprehensive data sets for Noida and Ghaziabad.

After collecting the raw data, the next step is data cleaning—a crucial phase in the data analysis pipeline. Cleaning involves several tasks to ensure the accuracy and consistency of the collected data. We handle missing values by either replacing them with appropriate values or removing them altogether. Additionally, we standardize data formats to ensure consistency across fields such as dates, prices, and addresses. Removing duplicate entries and correcting errors further enhances the quality of the dataset.

Week-6 Report

Date 11 Mar	From	To 16 Mar
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Descriptions:-

To facilitate data cleaning, we utilize Python libraries such as pandas, which offers robust tools for data manipulation and analysis. With pandas, we can load the scraped data into a structured format, perform cleaning operations efficiently, and store the cleaned data for further analysis.

It's important to note that while Selenium automates the data collection process, compliance with the terms of service of the government websites is paramount. Additionally, regular monitoring and updating of the code may be necessary to accommodate changes in the website's layout or structure.

In summary, leveraging the Selenium library enables us to automate the collection of property data from government websites for Noida and Ghaziabad. Through meticulous data cleaning, we ensure the accuracy and reliability of the collected data, laying the foundation for insightful analysis and informed decision-making in the real estate sector.

Week-7 Report

Date 18 Mar	From	To 23 Mar
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No.	Tasks Performed	Status
-------	-----------------	--------

Descriptions:— Comprehensive Data Analysis for Property Development: Insights into Certificate, Land Cost, and Construction Updates

In the intricate process of property development, navigating through a myriad of data points is essential for informed decision-making. From legal certifications to land costs and construction updates, each facet plays a pivotal role in shaping the trajectory of a project. This analysis delves into the backend data encompassing CA, architectural, and engineering certificates, land costs, and construction updates, offering invaluable insights for property developers and stakeholders alike.

Certifications from Chartered Accountants (CAs) serve as a testament to the financial viability and integrity of a property venture. These certificates meticulously examine financial records, ensuring compliance with regulatory standards and providing assurance to investors and financial institutions. Through data analysis, trends in CA certifications can unveil patterns of financial stability or identify areas requiring further scrutiny, empowering developers to make informed financial decisions.

Architectural and engineering certificates constitute the backbone of property development, encompassing structural integrity, design compliance, and adherence to building codes. Analyzing backend data related to these certifications offers insights into the quality and feasibility of a project. By scrutinizing historical data, developers can identify trends in design preferences,

technological advancements, and regulatory changes, thus optimizing future project planning and design strategies.

Land cost is a critical determinant in property development, exerting a profound influence on project feasibility and profitability. Backend data on land costs provides developers with a comprehensive understanding of market dynamics, including fluctuations in land prices, demand-supply dynamics, and regulatory interventions. By analyzing historical land cost data, developers can forecast future trends, assess investment risks, and strategically allocate resources, thereby maximizing returns on investment.

Construction updates encapsulate the dynamic progress of a property venture, reflecting milestones achieved, challenges encountered, and project timelines. Backend data on construction updates offers real-time insights into project progression, enabling developers to monitor performance, identify bottlenecks, and implement corrective measures promptly. Through data-driven analysis, developers can optimize construction schedules, mitigate delays, and ensure timely project delivery, thereby enhancing stakeholder satisfaction and project profitability.

Week-8 Report

Date 25 Mar	From	To 30 Mar
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Descriptions:— "Visualizing Application Movement Patterns Using Google Analytics
Data "

Understanding user behavior within an application is crucial for optimizing user experience and driving engagement. Leveraging data collected from Google Analytics, we can visualize user movement patterns within the application to gain insights into user interactions, navigation paths, and engagement levels. This report presents visualizations based on Google Analytics data to provide a comprehensive understanding of application movement.

Overview of Application Engagement:

- Pie chart depicting the distribution of user sessions by device type (mobile, desktop, tablet).
- Line chart illustrating the trend of daily active users (DAU) over time to identify patterns in user engagement.

1. User Navigation Paths:

- Sankey diagram showcasing the most common user navigation paths within the application.
- Funnel visualization highlighting the user journey from initial interaction to desired action (e.g., sign-up, purchase).

2. Page Interaction Analysis:

- Heatmap visualization displaying user clicks and interactions on various pages of the application.
- Bar chart showing the top pages visited by users, along with the corresponding number of sessions.

Week-9 Report

Date 1 Apr	From	To 6 Apr
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Ta	asks Performed	Status
----------	----------------	--------

User Engagement Metrics:

- Scatter plot depicting user engagement metrics such as session duration vs. bounce rate to identify trends and outliers.
- Histogram illustrating the distribution of session durations to understand user engagement levels.

Geographic User Distribution:

- Choropleth map visualizing the geographic distribution of application users across different regions or countries.
- Bar chart representing the top countries by user sessions, providing insights into global user demographics.

Event Tracking Analysis:

- Stacked area chart showcasing the frequency of specific events tracked within the application over time.
- Word cloud visualizing the most common event categories or actions performed by users within the application.

Week-10 Report

Date 8 Apr	From	To 13 Apr
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

1. Rechecking Data:

- Review the collected property data to ensure accuracy and completeness.
- Double-check for any discrepancies or errors that may have been missed during the initial data collection.

2. Cleaning Data:

- Clean the property data to remove inconsistencies, errors, and irrelevant information.
- Handle missing values by replacing them with appropriate values or removing them if necessary.
- Standardize data formats to ensure consistency across fields such as dates, prices, and addresses.
- Remove duplicates and correct any errors or inconsistencies in the data.

3. Replacing Values:

- Identify any specific values that need to be replaced based on predefined criteria or business rules.
- Replace the identified values with the correct or updated values.
- Ensure that the replacements are accurately reflected in the cleaned dataset.

4. Submitting Data to Backend Team:

• Prepare the cleaned and updated property data for submission to the backend team.

- Provide clear documentation or instructions on the data format and any changes made during the cleaning process.
- Collaborate with the backend team to facilitate the upload of the cleaned data into the application database.

5. Verification and Testing:

- Verify that the cleaned data has been successfully uploaded to the application backend.
- Test the functionality of the application to ensure that the updated property data is displayed correctly and accurately.
- Address any issues or discrepancies identified during testing and make necessary adjustments.

6. Making Properties Live:

- Coordinate with the backend team to initiate the process of making the properties live on the application.
- Monitor the application for any potential issues or errors that may arise after making the properties live.
- Communicate with stakeholders to notify them of the successful update and availability of the properties on the application.

7. **Documentation and Feedback:**

- Document the entire process, including data cleaning, replacement, submission, and making properties live.
- Gather feedback from stakeholders and team members to identify areas for improvement in the data collection and maintenance process.

Week-11 Report

Date 15 Apr	From	To 20 Apr
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Rechecking Data:

- Review the collected property data to ensure accuracy and completeness.
- Double-check for any discrepancies or errors that may have been missed during the initial data collection.

Cleaning Data:

- Clean the property data to remove inconsistencies, errors, and irrelevant information.
- Handle missing values by replacing them with appropriate values or removing them if necessary.
- Standardize data formats to ensure consistency across fields such as dates, prices, and addresses.
- Remove duplicates and correct any errors or inconsistencies in the data.

Week-12 Report

Date 22 Apr	From	To 27 Apr
Supervisor Name	Prakash Nagesh	
Designation	Senior associate	
Email	Prakash.nagesh@blackopalgroup.in	
Contact No.	9406988018	

S.No. Tasks Performed	Status
-----------------------	--------

Submitting Data to Backend Team:

- Prepare the cleaned and updated property data for submission to the backend team.
- Provide clear documentation or instructions on the data format and any changes made during the cleaning process.
- Collaborate with the backend team to facilitate the upload of the cleaned data into the application database.

Conclusion (Project Outcomes)

During my internship as a Market Research Analyst, I gained invaluable experience and developed a diverse skill set that has equipped me for a successful career in the field of data analysis and market research. This hands-on learning journey encompassed a wide range of technical and analytical skills, from data crawling and cleaning to practical applications in Excel and data manipulation. Additionally, I played a crucial role in providing backend data to enhance our application's functionality. Here's an in-depth look at the skills and experiences I acquired during this period.

Data Crawling Using Selenium

One of the most significant aspects of my internship was learning and implementing data crawling techniques using Selenium. Selenium, a powerful tool for web scraping and browser automation, allowed me to extract valuable data from various online sources. I learned to write scripts to navigate through websites, interact with web elements, and extract the necessary data efficiently. This skill is critical in market research, where up-to-date and comprehensive data is paramount for analysis and decision-making.

Data Cleaning

Raw data is often messy and unstructured, making data cleaning a crucial step in the data analysis process. During my internship, I became proficient in various data cleaning techniques to ensure the accuracy and quality of the data I worked with. I learned to identify and handle missing values, remove duplicates, standardize data formats, and correct inconsistencies. This process is essential to transform raw data into a reliable dataset that can be used for insightful analysis and reporting.

Practical Hands-on Experience with Excel

Excel remains a fundamental tool in data analysis, and my internship provided extensive hands-on experience with its powerful features. I became adept at using Excel for data organization, analysis, and visualization. Key skills I developed include using advanced functions and formulas, creating pivot tables, and generating various types of charts and graphs. Excel's ability to handle large datasets and perform complex calculations made it an invaluable tool for my tasks as a market research analyst.

Data Manipulation

Data manipulation involves transforming raw data into a usable format for analysis. During my internship, I gained experience in various data manipulation techniques using both Excel and programming languages such as Python. I learned to merge datasets, create new variables, filter and sort data, and perform complex data transformations. These skills are essential for preparing data for deeper analysis and ensuring that it meets the specific needs of our research projects.

Providing Backend Data for Application Enhancement

One of my key responsibilities was to provide accurate and timely data to support our application's functionality. This involved understanding the data requirements of our application, ensuring data integrity, and working closely with the development team. I developed an understanding of backend processes, including database management and data integration. By providing reliable data, I contributed to improving the application's performance and user experience, demonstrating the practical application of my data handling skills in a real-world context.

Bibliography

Reference for Selenium Webdriver:-

 $\underline{https://youtu.be/2DD\text{-}ynCIZ4w?si=eNZo_D0qUF5I2bgW}$



08th December 2023

Ankit Pawar

Vandana Vihar, Khora, Ghaziabad, Uttar Pradesh 201309

INTERNSHIP OFFER LETTER

Dear Ankit

Greetings from Blackopal Group! Welcome to the BOG Internship Program 2023.

We are pleased to have you on board as a Market Research intern this year. The internship will be for 3 months starting from 13th December 2023 to 14th March 2024. If the need arises, it could be extended further. During this period, you will be paid of stipend of Rs 10,000 per month.

The working hours are from 10:00 AM to 06:30 PM, Monday to Friday, with weekends off.

Your project details will be shared with you at the earliest. Your place of reporting will be our corporate office in Noida – Blackopal Group, 11th Floor, Gulshan One29, Sector 129, Noida 201304.

Wishing you all the best! We look forward to a mutually enriching experience.

Yours Sincerely,

Human Resources Blackopal Group

BLACK OPAL TECHNOLOGIES PRIVATE LIMITED

Gulshan One29, Plot no -C3 - E1 Sec-129, Noida-Greater Noida Expy, Noida, Ultar Pradesh 201301



blackopalgroup.in



12th March 2024

Ankit Pawar

Vandana Vihar, Khora, Ghaziabad, Uttar Pradesh 201309

INTERNSHIP EXTENSION LETTER

Dear Ankit

Greetings from Blackopal Group!

After reviewing your performance and considering the ongoing projects and opportunities within our company, we are pleased to offer you an extension of your internship on the following terms

The Internship stipend and other terms and conditions remain the same The duration would be from 13th March 2024 to 12th June 2024. During this period in case, you decide to resign, then we need a notice period of 1 week

Please consider this letter as formal notification of the extension of your internship.

Wishing you all the best! We look forward to a mutually enriching experience

Yours Sincerely.

Human Resources Blackopal Group

BLACKOPAL TECHNOLOGIES PRIVATE LIMITED



30th April 24

INTERNSHIP CERTIFICATE

Mr. Ankit Pawar has successfully completed his internship with BlackOpal Technologies (P) Ltd from December 13, 2023, to April 30, 2024. During his internship, he worked on the project "Data Collation of NCR Projects" in the Real Estate Segment.

We wish Ankit Pawar all the best in his future endeavors

Yours Sincerely,

Human Resburces **Blackopal Consultants**

BLACKOPAL TECHNOLOGIES PRIVATE LIMITED

Gulshan One29, 11th Floor, Sector 129, Noida, Uttar Pradesh-201304

Reg. Add.- Shop No. B, G Floor, C.S.C, AGCR Enclave, Delhi-110092 | GST No.- 07AAKCB1310K1ZX, CIN No.- U74300DL2021PTC386413



