Internship Report

Automated Tracking of eCommerce Levers for Amazon

Submitted by: Tushya Gupta
Roll number: UE233106

Acknowledgement

I would like to express my sincere gratitude to my college, *University Institute of Engineering and Technology*, for providing me with the opportunity to undertake this internship. I am deeply thankful to the Head of Department for his constant encouragement and support. I extend my appreciation to my faculty guide for their valuable guidance throughout the internship. My heartfelt thanks to the professors and teaching staff for the knowledge and skills imparted during my course. I am grateful to the administrative staff for their timely help and cooperation. Special thanks to my friends and classmates for their continuous motivation and assistance. I also thank the training and placement cell for facilitating the internship process. This internship would not have been possible without the strong academic foundation provided by my college. I sincerely acknowledge the support of everyone who contributed directly or indirectly to this work. Finally, I remain indebted to my institute for shaping my learning journey and career.

Executive Summary

The internship focused on building automated scrapers, data pipelines, and visualization tools to analyse product performance on Amazon. The objectives included tracking product rankings across regions, monitoring BuyBox ownership, studying delivery timelines, and gaining insights into order statuses and returns. Over seven weeks, multiple scrapers were developed using different Python libraries, supported by proxies and browser impersonation to bypass detection. These tools were modular, fault-tolerant, and designed with backup and retry mechanisms for reliable data collection.

Collected data was integrated with Google Sheets APIs and visualized using Looker Studio dashboards. Delivery timelines, product rankings, BuyBox ownership, and competitor presence were mapped geographically, highlighting strong performance in metropolitan areas but slower deliveries in Tier-2 and Tier-3 cities. The analysis also revealed competitor dominance for certain keywords and the entry of new sellers offering similar products. Order status and return tracking provided operational insights into fulfilment bottlenecks and products with high return rates.

The project demonstrated the value of data-driven approaches in e-commerce. By automating monitoring and using interactive dashboards, the company gained actionable insights to optimize logistics, strengthen BuyBox control, improve product visibility, and reduce returns. Despite challenges such as CAPTCHAs, proxy reliability, and storage limitations, the tools created showcased the potential of scalable scraping and visualization for strategic improvements.

Table of Contents

Company Profile

Everything Beautiful Retail is an e-commerce company engaged in selling a wide range of consumer products through leading online platforms, with Amazon being its primary marketplace. The company is committed to providing high-quality products, ensuring customer satisfaction, and maintaining efficient delivery services. In addition to its main retail operations, it also manages a brand called *BoxJoy*, which offers curated product bundles and unique offerings designed to enhance customer experience and convenience.

By combining strong retail expertise with digital marketplace strategies, Everything Beautiful Retail has steadily grown its online presence and broadened its customer base across India. The company continuously explores innovative methods to improve product visibility, strengthen brand identity, and adapt to the evolving dynamics of e-commerce. Its focus on data-driven decision-making and technology adoption allows it to stay competitive while delivering value to customers.

Introduction

During my internship, the primary focus was to address challenges faced by online sellers in monitoring and improving their product performance on Amazon. In the highly competitive e-commerce space, visibility of products, ownership of the BuyBox, and delivery efficiency are key factors that directly affect sales. Sellers often lack precise tools to track how their products rank across different regions or pincodes, which sellers control the BuyBox, or how delivery timelines vary from state to state. Without this data, it becomes difficult to evaluate performance against competitors, optimize listings, or improve supply chain strategies.

My work was cantered on designing automated scrapers and analytical tools that could collect, process, and present this information in a structured and actionable format. This involved building systems to fetch product rankings based on keywords across multiple regions, identifying the current BuyBox seller along with their pricing and ratings, and tracking delivery timelines of both our products and competitors. The data was further integrated into Google Sheets and visual dashboards, enabling quick interpretation and comparison across markets. In addition, solutions were developed to overcome real challenges such as Amazon's CAPTCHA checks, proxy blocking, large data volumes, and session handling.

By the end of the internship, the outcome was a set of automated tools and reporting systems that provided the company with deeper visibility into product performance on Amazon. This work not only streamlined data collection but also offered actionable insights that could support strategic decision-making in sales and operations.

Research Methodology

The research methodology adopted during this internship was primarily technical and exploratory, aimed at developing automated tools to extract, process, and analyse product performance data from Amazon. Since Amazon's platform does not provide this information openly, a structured approach was required to design scrapers that could reliably capture ranking, seller, and delivery data across multiple regions in India.

The first step involved building scrapers using Python with libraries such as Selenium, httpx, curl_cffi, and Parsel for HTML parsing. Different approaches were tested, starting with browser automation and later transitioning to non-browser methods for improved speed and stealth. Proxy rotation, TOR networks, and browser impersonation were used to bypass detection and reduce the chances of blocking. A modular coding structure was followed, enabling independent development of ranking scrapers, BuyBox checkers, and delivery timeline scrapers. Error handling mechanisms such as retries, checkpoints, and backup systems ensured data collection could continue even after failures.

To manage and visualize the large volumes of data, integration with the Google Sheets API was implemented, along with CSV exports for long-term storage. Authentication through OAuth allowed secure access and eliminated reliance on temporary scripts. Looker Studio dashboards were then created to map delivery times, seller presence, and ranking distribution across states and pincodes. Additional layers of analysis included comparative timelines with top competitors, BuyBox price differences, and keyword search volume estimation using external reports and Amazon's SP-API.

This methodology combined web scraping, data engineering, and visualization techniques to create a pipeline that was not only functional but scalable and adaptable to changing conditions. By iteratively improving the tools each week and validating the outputs against known benchmarks, the project ensured that the insights generated were both accurate and actionable.

Data Analysis and Interpretation

The data collected during the internship was cantered around three major areas: product rankings, BuyBox ownership, and delivery timelines. Each dataset provided unique insights into how products performed on Amazon, both individually and in comparison to competitors. Since the e-commerce marketplace is highly dynamic, capturing, cleaning, and interpreting this information was critical to ensuring its usefulness.

1. Product Rankings Across Regions
The ranking scrapers provided data on how products appeared in search results for given keywords across different pincodes in India. By mapping these results to states and cities, it became possible to see regional differences in visibility. For example, certain products ranked highly in metropolitan areas but performed poorly in smaller towns. This pattern indicated that regional stock availability, demand, and competition played a significant role in search placement. The ranking data was exported to CSV and further visualized in Looker Studio maps, where each state and region was color-coded according to ranking performance. This visualization simplified the interpretation process, making it easier to identify consistent strongholds and weak spots across India.

2. BuyBox Ownership and Seller Analysis
The BuyBox is a crucial factor in Amazon sales, as it determines which seller gets the majority of purchases for a given product. The BuyBox checker scraped details such as the current seller, their pricing, fulfilment method (FBA/MFN), and additional offers available. By comparing BuyBox results across time and regions, we were able to detect trends such as price competition, temporary ownership changes, and promotional impacts. The integration with Looker Studio allowed BuyBox presence to be plotted geographically, showing where the company dominated and where competitors held control. This geographic interpretation of BuyBox data provided a clear picture of regional competition and pricing strategies.

3. Delivery Timeline and Service Coverage

Delivery time is another key driver of customer satisfaction and ranking performance. By scraping delivery timelines for both the company's products and the top five organically ranked competitors, we were able to compare service quality. Looker Studio maps were used to highlight delivery speed across India, with faster deliveries represented in green and slower deliveries in red. This visualization provided immediate insight into regions where the company's logistics were strong and where improvements were needed. For large states, the map also revealed significant urban-rural delivery gaps, guiding potential logistics interventions.

4. Search Volume and CTR Estimation

To complement ranking and delivery data, keyword search volume was obtained through Amazon SP-API reports and external tools. This data was integrated with ranking results to estimate potential click-through rates (CTR) for products. For example, a product ranking in the top three positions for a high-volume keyword had significantly greater visibility compared to one ranking lower for the same keyword. Weighted CTR values across pincodes were then calculated and plotted on Looker Studio dashboards, allowing the company to visually assess opportunity hotspots by region and keyword.
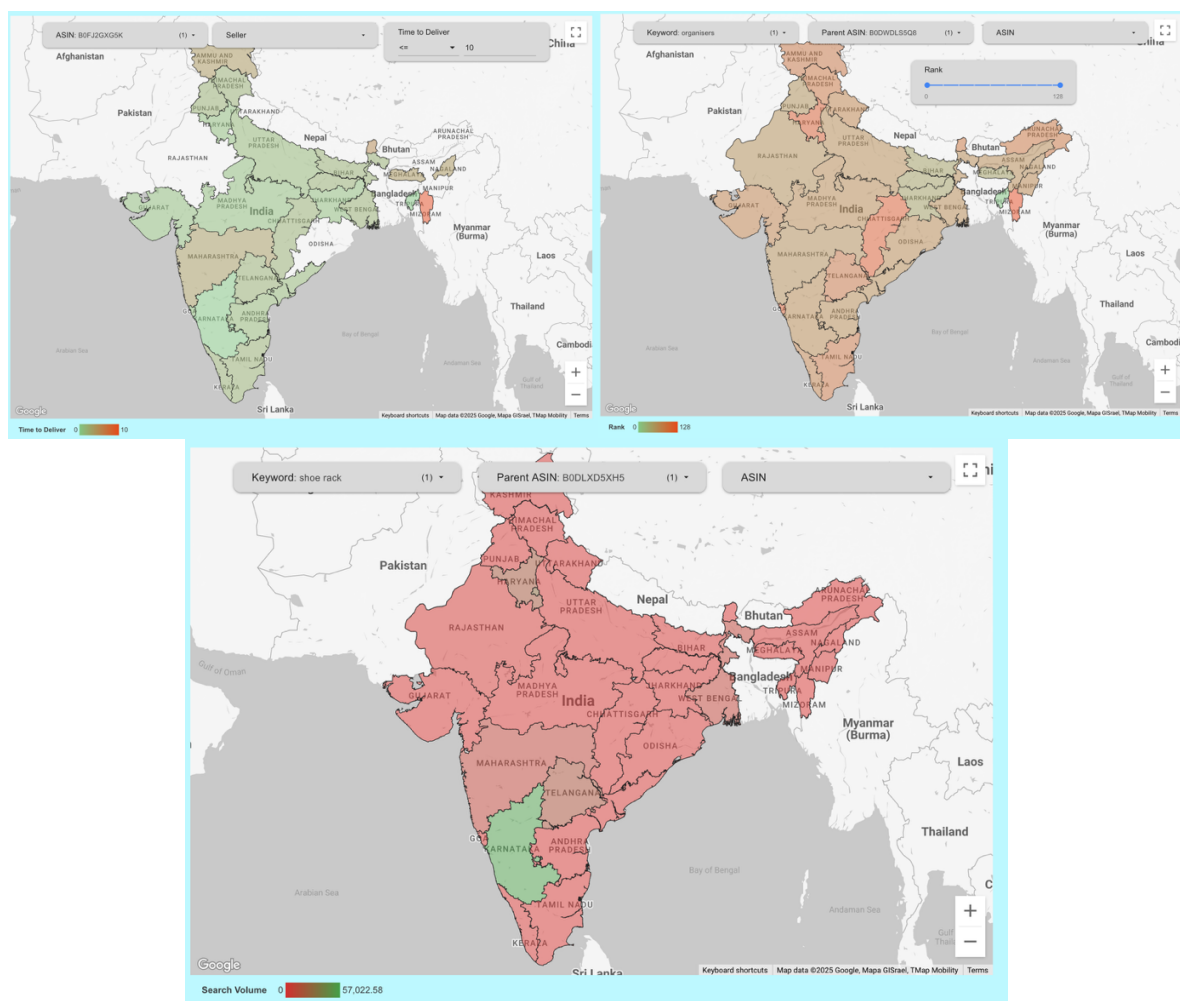
5. Order and Return Analysis

In addition to front-end product data, backend reports from Amazon SP-API provided insights into order statuses and returns. By analysing purchase order fulfil ability and identifying products with higher return rates, we could pinpoint operational issues. This data was also represented in Looker Studio to give a regional distribution of return rates and fulfillable orders, making operational bottlenecks more visible. The financial interpretation of this analysis allowed the company to estimate revenue at risk due to stockouts or return trends.

Interpretation and Insights

The combined datasets, when visualized through Looker Studio maps and dashboards, revealed several important patterns. Firstly, product performance was not uniform across India; regional differences in ranking and delivery times highlighted the need for tailored logistics strategies. Secondly, BuyBox ownership was closely tied to both pricing and fulfilment methods, indicating that a balanced strategy between competitive pricing and fast delivery was essential. Thirdly, the visualization layer greatly simplified interpretation, turning large datasets into interactive and easy-to-understand maps that stakeholders could use for decision-making.

Overall, the analysis showed that the company's strength lay in competitive delivery timelines in certain urban areas, while weaknesses included inconsistent BuyBox ownership and ranking variability in smaller markets. The use of Looker Studio maps added a powerful dimension to this analysis by making insights more accessible, interactive, and actionable.

Findings

The systematic collection and analysis of Amazon product data through scrapers, APIs, and visualization tools generated several significant findings across ranking, delivery, competition, and order management.

1. Regional Delivery Disparities

The delivery timeline analysis showed that while metro regions like Delhi, Mumbai, and Bangalore enjoyed quick delivery times, several Tier-2 and Tier-3 cities experienced longer delays. These delays were linked to warehouse proximity and courier availability. Inconsistent delivery performance was also observed within the same state, highlighting gaps in regional logistics networks. Faster delivery times correlated with higher ranking stability, confirming that delivery speed plays a role in organic visibility.

2. Competitor Dominance in Rankings

Ranking analysis revealed that competitors had overtaken the company's products in certain regions and for specific keywords. In high-volume categories, competitor listings frequently appeared as the top result, reducing the visibility of the company's products. This finding implies that product ranking is not just influenced by keyword optimization but also by delivery performance, BuyBox control, and possibly advertising investments.

3. Emergence of New Sellers and Market Saturation

The BuyBox scraper provided evidence that new sellers had entered the marketplace with similar products. These sellers often competed aggressively on price, with some leveraging Fulfilment by Amazon (FBA) to gain an advantage in delivery speed and trustworthiness. This trend indicates an increase in market saturation and suggests that differentiation through branding, better product descriptions, or bundled offerings may be necessary to maintain market share.

4. BuyBox Ownership Trends

Detailed BuyBox tracking showed that ownership shifted between sellers more frequently than expected. In some cases, ownership alternated within a short time frame due to price changes or stockouts. This volatility demonstrates that BuyBox control is dynamic and requires continuous monitoring. Products where the company lost BuyBox ownership also showed a corresponding dip in sales velocity, reinforcing its critical role in driving conversions.

5. Order Status and Returns

The order status scripts revealed valuable insights into the fulfilment process. By tracking order progression from placement to delivery, it was possible to identify bottlenecks. The return analysis highlighted certain products with above-average return rates, often linked to issues such as product quality mismatches, packaging damage, or unrealistic delivery promises. Returns not only caused direct revenue loss but also increased operational costs, suggesting the need for targeted corrective measures.

6. Search Volume and Opportunity Sizing

Integrating keyword search volumes with ranking and CTR estimates highlighted untapped opportunities. Certain high-volume keywords showed low product visibility for the company, indicating potential growth areas. Conversely, in some lower-volume niches, the company already ranked well, which could be leveraged to strengthen its dominance without heavy competition. This balance between opportunity and competitiveness provides a roadmap for keyword targeting and campaign focus.

7. Impact of Data Visualization

Using Looker Studio maps and dashboards significantly improved the interpretation of these findings. Visual layering of rankings, delivery times, and seller distribution allowed stakeholders to quickly spot problem areas. For example, a heatmap view immediately highlighted slow delivery clusters in northern and northeastern states, while BuyBox maps identified competitor strongholds in western India. This visualization-driven insight made it easier to translate raw data into actionable strategies.

Interpretation of Findings

The findings collectively demonstrate that the company's performance on Amazon is shaped by a combination of operational efficiency, competitive positioning, and product visibility. Strengths include consistent delivery times in metropolitan areas and the ability to build automated monitoring systems. Weaknesses are visible in regional delivery inconsistencies, volatile BuyBox ownership, and rising competition from new entrants. At the same time, keyword opportunity analysis has uncovered growth areas where the company can strengthen visibility and sales.

The data-driven approach not only diagnosed current challenges but also highlighted opportunities for future growth. These findings form the foundation for strategic recommendations in areas such as logistics optimization, pricing strategy, product differentiation, and proactive BuyBox monitoring.

Suggestions

Based on the analysis and interpretation of the data, several suggestions can be made to improve product visibility, strengthen competitiveness, and optimize operations for the company. These recommendations focus on addressing delivery gaps, enhancing BuyBox ownership, and improving long-term efficiency.

1. Strengthen Regional Logistics
Since delivery speed was observed to directly impact both rankings and customer satisfaction, efforts should be made to optimize logistics in underperforming regions. This can include stocking products more strategically across fulfilment centres, partnering with reliable regional couriers, or expanding the use of Amazon FBA in areas where delivery times are consistently slower. Reducing delivery delays in Tier-2 and Tier-3 cities would improve competitiveness and increase the likelihood of better rankings.

2. Targeted Keyword Optimization
The ranking data combined with search volume analysis shows that some high-volume keywords present untapped opportunities. The company should focus on improving listing content, keyword placement, and advertising strategies for these terms. At the same time, niches where the company already ranks well should be further reinforced to maintain dominance with relatively low effort. A balanced keyword strategy will improve both visibility and conversion rates.

3. Proactive BuyBox Monitoring and Pricing Strategy
BuyBox ownership volatility suggests the need for continuous monitoring and dynamic pricing adjustments. Automating BuyBox tracking to issue alerts when ownership is lost would enable quicker responses. Implementing competitive but sustainable pricing strategies, while leveraging FBA for faster deliveries, can help regain and maintain BuyBox control. In addition, bundling products or offering value-added services could reduce direct price competition.

4. Product Quality and Return Reduction

Return analysis showed certain products with higher-than-average return rates. A focused quality review should be conducted for these items, addressing issues like packaging durability, accuracy of product descriptions, or defects in the product itself. Reducing return rates will not only increase revenue but also improve customer satisfaction and seller ratings, indirectly boosting product visibility.

5. Competitor Analysis and Differentiation

The entry of new sellers into the same product categories highlights the need for stronger product differentiation. The company could enhance listings with better images, detailed descriptions, and value-based messaging. Developing unique bundles under the *BoxJoy* brand can also differentiate offerings and make it harder for competitors to directly replace products.

6. Operational Dashboards and Decision Support

The use of Looker Studio dashboards proved valuable during the internship. It is recommended that the company continue to expand these dashboards into a real-time monitoring system that can be used by sales and operations teams. Automating updates and including performance alerts would ensure that decision-making remains data-driven and timely.

Summary of Suggestions

Improving delivery times in slower regions, refining keyword strategies, actively monitoring BuyBox control, reducing returns, differentiating products, and expanding dashboard usage are key steps that will strengthen the company's position on Amazon. Implementing these measures will enable Everything Beautiful Retail to improve competitiveness, grow sales, and sustain long-term success in the marketplace.

Limitations

Despite the successful development of scrapers, dashboards, and automation tools, the project faced several limitations that restricted its scope and efficiency.

1. CAPTCHAs and Bot Detection by Amazon
One of the biggest challenges was Amazon's active bot detection systems. Scrapers frequently encountered CAPTCHAs, throttling, or blocking, which disrupted continuous data collection. While solutions like retries, proxy rotation, and browser impersonation were implemented, these only reduced the frequency of interruptions rather than eliminating them entirely. This meant that very large-scale or long-duration scraping runs were difficult to execute reliably.

2. Proxy Dependence and Reliability Issues
The use of rotating proxies and TOR-based connections was critical for bypassing detection but introduced its own limitations. Many proxies were either slow, already blacklisted by Amazon, or unstable during extended scraping sessions. Procuring reliable proxies also introduced additional costs. The constant need to validate and maintain a pool of functional proxies increased the overhead of running the scrapers effectively.

3. Dynamic and Unpredictable Website Structures
Amazon frequently modifies its HTML structures, session management methods, and backend APIs. Even minor changes in the layout or request handling often broke parsing logic, requiring manual intervention to update the scrapers. This lack of structural consistency created ongoing maintenance challenges and limited the robustness of the system. It also meant that some data fields were occasionally missed or incorrectly parsed.

4. Limitations of Data Storage and Visualization Tools

Although Google Sheets and Looker Studio proved effective for smaller datasets, they were not ideal for managing very large volumes of data. Google Sheets has a cell limit (10 million), and once this threshold was approached, performance slowed significantly. CSV exports solved part of this issue, but they lacked real-time collaboration features and interactivity. Similarly, while Looker Studio provided powerful visualizations, handling large or frequently updated datasets sometimes caused performance bottlenecks.

5. Dependence on Third-Party Data and APIs

Certain metrics, such as keyword search volumes and competitive benchmarks, were derived from Amazon's SP-API and external tools. The accuracy of these results depended heavily on the reliability of the third-party sources. In cases where data was incomplete, delayed, or unavailable, the analysis was limited in depth. Furthermore, reliance on external APIs introduced potential risks if those services changed their access policies or became unavailable in the future.

Summary of Limitations

In summary, the project faced constraints due to anti-bot measures, proxy dependencies, constantly evolving website structures, storage and visualization limitations, and reliance on external data sources. While workarounds were developed for most challenges, these limitations highlight the inherent difficulties of large-scale data scraping and analysis in a dynamic e-commerce environment.

Conclusion

The internship provided a comprehensive learning experience in the fields of web scraping, automation, and data-driven analysis for e-commerce. The central objective was to design and implement systems capable of monitoring Amazon product rankings, BuyBox ownership, delivery timelines, and order statuses across different regions of India. Over the seven-week period, multiple scrapers and analytical tools were developed, tested, and refined, each addressing a specific challenge faced by the company.

The work resulted in several tangible outcomes. Automated scrapers were successfully built to track rankings and BuyBox sellers, while integration with Google Sheets and Looker Studio allowed the visualization of delivery times, regional product performance, and competitor activity. Order status and return analysis scripts provided operational visibility into customer orders and highlighted areas of improvement in product quality and logistics. The findings showed clear patterns—delivery disparities across regions, competitor dominance in certain markets, and rising competition from new sellers—all of which offered practical insights for strategic decision-making.

At the same time, the project highlighted key limitations. CAPTCHAs, proxy reliability issues, dynamic website structures, and storage constraints presented challenges that required creative workarounds. The dependence on third-party APIs and external tools further emphasized the need for caution when scaling such systems. Despite these challenges, the methodology proved effective, and the outcomes demonstrated the value of combining scraping, automation, and visualization for actionable insights.

Overall, the internship strengthened technical proficiency in Python, APIs, and data visualization while also improving problem-solving, documentation, and project management skills. More importantly, it provided the company with tools and insights that can guide improvements in logistics, keyword optimization, pricing, and product differentiation. The project reaffirmed the importance of data-driven strategies in e-commerce and showcased how even small-scale technical interventions can lead to meaningful business outcomes.