

|  |
| --- |
|  |
| WALMART 2.0 |
| ETL PROJECT |
| AWESOME POSSUM | [github](https://github.com/beckyydo/Project-2) |



|  |  |
| --- | --- |
|  | Decorative |
| contents |
| [**03**](#_Introduction)  INTRODUCTION  [**04**](#_methodology)  METHODOLOGY  [**06**](#_results)  RESULTS  [**08**](#_WHAT’S_NEXT)  WHAT’S NEXT |

|  |  |
| --- | --- |
| Group brainstormGROUP MEMBERS  * Hanieh Babaee * Ronald Clarke * May Lacdao * Rebecca Pham * Elizabeth Salas-Martinez | Decorative |
|  | Introduction |
|  | motivation In business, gaining a competitive edge entails an understanding of the business' progress and the factors that affect it. Current technological advancements allow businesses to create business value through data analytics. Retail giants, like Walmart and Amazon, amass huge amounts of data, which, when processed with the right set of tools could provide powerful actionable insights.  For this ETL project, our group decided to expand on our first project and continue working on Walmart datasets. data sources Data sets were retrieved from the following sources:   * Kaggle * data.world * Reddit * pypi * Institute for Local Self-Reliance |
|  |  |
|  |  |

|  |  |  |
| --- | --- | --- |
|  |  | Decorative |
|  |  | methodology |
| extraction Lorem ipsum dolor sit amet consectetur adipiscing. Nisi lacus sed viverra tellus. Orci eu lobortis elementum nibh tellus molestie nunc non. Laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Pharetra massa ultricies mi quis hendrerit dolor. Non tellus orci ac auctor augue mauris augue neque gravida. transformation Lorem ipsum dolor sit amet consectetur adipiscing. Nisi lacus sed viverra tellus. Orci eu lobortis elementum nibh tellus molestie nunc non. Laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Pharetra massa ultricies mi quis hendrerit dolor. Non tellus orci ac auctor augue mauris augue neque gravida.   loading Lorem ipsum dolor sit amet consectetur adipiscing. Nisi lacus sed viverra tellus. Orci eu lobortis elementum nibh tellus molestie nunc non. Laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Pharetra massa ultricies mi quis hendrerit dolor. Non tellus orci ac auctor augue mauris augue neque gravida.  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Quis eleifend quam adipiscing vitae proin sagittis nisl rhoncus mattis. At tellus at urna condimentum mattis. Odio aenean sed adipiscing diam donec adipiscing tristique. Molestie ac feugiat sed lectus vestibulum mattis ullamcorper velit sed. Cursus in hac habitasse platea dictumst quisque sagittis purus sit. Cursus sit amet dictum sit amet justo donec |

|  |  |  |
| --- | --- | --- |
| Decorative |  |  |
| methodology |  |  |
| extraction Lorem ipsum dolor sit amet consectetur adipiscing. Nisi lacus sed viverra tellus. Orci eu lobortis elementum nibh tellus molestie nunc non. Laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Pharetra massa ultricies mi quis hendrerit dolor. Non tellus orci ac auctor augue mauris augue neque gravida. transformation Lorem ipsum dolor sit amet consectetur adipiscing. Nisi lacus sed viverra tellus. Orci eu lobortis elementum nibh tellus molestie nunc non. Laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Pharetra massa ultricies mi quis hendrerit dolor. Non tellus orci ac auctor augue mauris augue neque gravida.   loading Lorem ipsum dolor sit amet consectetur adipiscing. Nisi lacus sed viverra tellus. Orci eu lobortis elementum nibh tellus molestie nunc non. Laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Pharetra massa ultricies mi quis hendrerit dolor. Non tellus orci ac auctor augue mauris augue neque gravida.  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Quis eleifend quam adipiscing vitae proin sagittis nisl rhoncus mattis. At tellus at urna condimentum mattis. Odio aenean sed adipiscing diam donec adipiscing tristique. Molestie ac feugiat sed lectus vestibulum mattis ullamcorper velit sed. Cursus in hac habitasse platea dictumst quisque sagittis purus sit. Cursus sit amet dictum sit amet justo donec |  |  |

|  |
| --- |
| Decorative |
| results |
| Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Feugiat vivamus at augue eget. Vitae turpis massa sed elementum tempus. Bibendum enim facilisis gravida neque. Aenean euismod elementum nisi quis eleifend quam adipiscing vitae proin. Facilisi nullam vehicula ipsum a. Integer enim neque volutpat ac tincidunt vitae semper. Proin sagittis nisl rhoncus mattis rhoncus. Morbi tristique senectus et netus. Purus ut faucibus pulvinar elementum integer.  Ut ornare lectus sit amet est placerat. Tincidunt vitae semper quis lectus nulla at volutpat. Sit amet nisl suscipit adipiscing bibendum. Nunc sed id semper risus in hendrerit. Tincidunt praesent semper feugiat nibh sed pulvinar proin gravida. Sed felis eget velit aliquet sagittis. Porta lorem mollis aliquam ut porttitor leo a diam. Mattis molestie a iaculis at erat pellentesque adipiscing commodo. |
|  |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Q1 | Q2 | Q3 | Q4 | | SECTOR 1 | 4.3 | 2.5 | 3.5 | 4.5 | | SECTOR 2 | 2.4 | 4.4 | 1.8 | 2.8 | | SECTOR 3 | 2 | 2 | 3 | 5 | |
|  |
|  |

|  |  |
| --- | --- |
|  | Decorative |
|  | WHAT’S NEXT |
|  | challenges Datasets and Data sources   * Walmart Sales 2010-2012: This dataset only included location details for Walmart stores. It also included information about other store types unrelated to Walmart. * Sales Zip File: This dataset contained several unknowns and unnamed variables. The format of the data required extensive manipulation to create columns by day for sales data, making the data extraction difficult. Despite the cleaning process, the dataset was still over 500MB and over 18 million rows. * Financial Data re. Walmart, Amazon, Target and COSTCO: A more thorough month-by-month or week-by-week breakdown of the stats would be more helpful in providing more granular tracking of each corporation's trajectory. This would also afford more opportunities to aggregate specific data for more focused ranges of time. * Market Share Data by Region: Pre-processing was required to obtain the Walmart Market Share data. The data is originally in pdf and jpg format and needed OCR technology, to retrieve information from jpg and save it in .csv format. * Online Product Database: The e-commerce dataset has 30,000 rows and appears to be a sample from a larger file. We decided to only use the first level of the e-commerce category, since the price/sales data are categorized at a similar level.   ETL Process   * Storage and Database creation: Too many tables were generated, leading to storage issues * Github: Intermediary files were added to .gitignore to free up space |

|  |  |
| --- | --- |
| Decorative |  |
| WHAT’S NEXT |  |
| future goals Datasets and Data sources   * Walmart Sales 2010-2012: This dataset only included location details for Walmart stores. It also included information about other store types unrelated to Walmart. * Sales Zip File: This dataset contained several unknowns and unnamed variables. The format of the data required extensive manipulation to create columns by day for sales data, making the data extraction difficult. Despite the cleaning process, the dataset was still over 500MB and over 18 million rows. * Financial Data re. Walmart, Amazon, Target and COSTCO: A more thorough month-by-month or week-by-week breakdown of the stats would be more helpful in providing more granular tracking of each corporation's trajectory. This would also afford more opportunities to aggregate specific data for more focused ranges of time. * Market Share Data by Region: Pre-processing was required to obtain the Walmart Market Share data. The data is originally in pdf and jpg format and needed OCR technology, to retrieve information from jpg and save it in .csv format. * Online Product Database: The e-commerce dataset has 30,000 rows and appears to be a sample from a larger file. We decided to only use the first level of the e-commerce category, since the price/sales data are categorized at a similar level.   ETL Process |  |