Creation and Development of a data extraction package in RStudio for improving the access to the open data of city of Windsor and its subsequent usage to perform EDA and subsequent analysis

Open data is one of the best solutions for promoting public oversight, public service improvement and innovation.(Government et al., 2019) This data is present in different formats categorised by individual issues on the City of Windsor open data portal and accessing this data becomes very difficult if a user or users want to perform comprehensive analysis on more than one public affairs. This is because the user would have to download each and every file from one or more public affairs' individual web pages separately onto the system and then go about studying the data individually. This method would never work for comparing response times between different affairs and performing overall comprehensive analysis. At the same time, this method is extremely time consuming and inconvenient. **opendatawindsor** package will enable the users to use the excel and csv data on the City of Windsor open data portal directly in their RStudio without the need to download those multiple individual files on their systems. This will enable the users to perform analysis on the data quickly and efficiently. The users of this package are not limited to data analysts/scientists but also the city officials who can understand how to better service the public.

Analytical Goals

To perform general EDA and develop a dashboard based on the analysis of specific public affairs from the list of datasets after using the package to load the datasets into R directly.

Existing Solutions for the problem/Packages

1. **opendatatoronto**(Open & Portal, n.d.) is a package that acts as an R interface for the City of Toronto Open Data portal. This package is the main motivation for our package and works as an excellent guide as it is developed for a similar purpose as it allows users to load the data into R directly without downloading the files on the system. The **difference** is that while this package will allow the users to load any and all datasets available on the City of Toronto open data portal, while **opendatawindsor** will allow the users to load only the excel and csv files from the open data portal of City of Windsor.

2. **LOPDF** (Linked Open Publications Data Framework): (Aslam, 2021) This is a generic framework for extracting and producing open data of scientific documents for smart digital libraries. This **framework** has been developed specifically for enabling the availability of the **metadata** of the scientific documents in a form which would enable the usage of smart queries that can help in computing and performing different types of analysis on scientific publications data. Also, the time and resources required for the acquisition and smart processing of publications data will be reduced and the task will be simplified by using this framework. The **difference** between this framework and our package is that this framework is specifically designed to work on the

metadata of scientific documents available on open scientific forums, websites, etc, while our package will extract the main datasets from the websites into R directly without having to download the data onto their individual systems.

Skills and Tools Required for the successful completion of this project

1. Knowledge of Package Creation and execution in RStudio.
2. Data Analysis RStudio capabilities for EDA after the package is run successfully.
3. Successful and smooth collaboration between all team members.
4. Swift and efficient problem solving/issue resolution.
5. Tableau Desktop for dashboard creation and analysis.
6. MS PowerPoint and Word for documenting our work and presenting/reporting it.
7. GitHub and its successful navigation and publishing skills will be necessary for the successful documentation of the package as well as project uploading for our individual portfolios.

Possible Challenges

1. We are going to try and create our package without using tidyverse's purrr function which would include us defining our own functions to create the package.
2. This is the first time we are developing our own package from scratch and we have always used R's existing packages and functions directly only for analysis which is why this is an unprecedented challenge.
3. Developing the package is not the only task in our project, there might be time issues later as the package development may take longer than necessary which would reduce the time available for EDA and dashboard creation.
4. Publishing and releasing the package will also be a new challenge as we need to make sure that every external/internal resource is credited appropriately.

Ethical Concerns

Since all of the data is open data available on the City of Windsor, the ethical concerns are minimal.

Any data which is personally identifiable is appropriately anonymized by the city itself.

Along with that, we will make sure that we use the data responsibly. We can only hope that the users of the opendatawindsor package make sure that they use any of the data they use responsibly as well.

Data Source

City of Windsor Open Data portal: https://opendata.citywindsor.ca/

Definition of Success/MVS (Minimum Viable Solution)

1. Development and successful execution of a working package with the least dependencies to extract data from the City of Windsor Open Data portal.
2. Once package development is complete, we will perform EDA and create a dashboard which will display the analysis for some of the public affairs from a total of 96 affairs.

Extenuating Condition:

We are going to try our best to develop the package without using tidyverse but should that not happen in the stipulated time, we will use the purrr function in tidyverse to extract the data from the websites for the appropriate function and hence the package development will be slightly modified.

Overwhelming Success

If our package is uploaded on CRAN, then that will make this project a grand success.

Evaluation Metrics

1. Minimum issues reported on our GitHub package repository with no crash or other errors when used in straightforward manner.

2. If our package is uploaded on CRAN, then we are expecting minimum errors or issues reported with the usage of the package.

Basic Pipeline

i) Development of **opendatawindsor** package.

ii) Selection of Issues(Public affairs) with the help of our project guide.

iii) Use of the package to extract the datasets of only these issues(public affairs).

iii) EDA on these selected issues(Public affairs).

iv) Final analysis and dashboard creation in Tableau Desktop along with publishing of this dashboard.

Collaborators

Chintan Wagh - https://github.com/chintanwagh

Devashish Patel - https://github.com/devashish-patel47

Jervis Murzello - https://github.com/JervisMurzello

Omkar Shinde - https://github.com/omkarshinde2798

Urvashi Ketulkumar Prajapati - https://github.com/URV45H1

References

1) Government, O., Toolkit, D., Essentials, O. D., Policies, O. D., Data, O., Resources, L., Options, T., & Tool, R. A. (2019). Starting an Open Data Initiative | Data. *The World Bank*, 1–9. http://opendatatoolkit.worldbank.org/en/starting.html

2) Open, T., & Portal, D. (n.d.). *opendatatoronto*. 1–5.

3) Aslam, M. A. (2021). LOPDF: A framework for extracting and producing open data of scientific documents for smart digital libraries. *PeerJ Computer Science*, *7*, 1–23. https://doi.org/10.7717/PEERJ-CS.445