6in6: Project Standards Proposal

UofT SCS Data Analytics

Project 2 – Group 6

2019 January 16

# File Naming Standards

In order not to accidentally affect code from other members of the group, I am proposing that we follow the following file naming conventions for our python notebook files.

All filenames will start with prefix 6in6\_

All notebook filenames will include what the activity the code will perform and what sources, camelback format with underscores separating ideas:

Therefore, we can work independently, and then be able to assemble the code easily at the end.

Similarly, we should use the same ideas for graphs and csv file products but with very specific titles.

For the final code, we will need to discuss what files should be executed first. I propose to add a two digit number to append to the 6in6\_prefix so that it will become clear what code to execute in which order.

Eg. 6in6\_01\_ExtractTransform\_IndeedJobs.ipynb

6in6\_02\_ExtractTransform\_GlassDoorJobs.ipynb

Eg. of various file types:

|  |  |  |
| --- | --- | --- |
| File Type | Filename Example | Description |
| Jupyter Notebook Interim | 6in6\_ExtractTransform\_IndeedJobs.ipynb | The Jupyter notebook that will extract data from the Indeed website and transform and save to the Jobs table, inbterim code |
| Graphs | 6in6\_BarGraph\_AverageDataScientistSalary     VAverageHousingPrice.png | The Bar Graph created showing the relationship between Average Data Scientist Salary and Average Housing Prices per provinces |
| Final Report | 6in6\_DataAnalyticJobsInTheCanadianMarket.doc | Final Report (this name is a suggestion) |
| Jupyter Notebook Final | 6in6\_01\_ExtractTransform\_IndeedJobs.ipynb  6in6\_02\_ExtractTransform\_GlassDoorJobs.ipynb | Final code showing the sequence to run |

I also propose that we have a living Data Dictionary document that we all update and use so that we have something to refer to as we code: This includes schema, table, and column names.

One of the things I like to do is provide an appendix in the final report of all the files we have as in the table above

Versioning: We should add \_v01 as we version different files.

# File Folder Names

The following is the folder structure:

|  |  |
| --- | --- |
| Folder | Description |
| 0\_Reference | This will include any reference documents including the original project proposal & this Standards proposal, and any pictures and images (non generated) we may want to include in our Final Report |
| 1\_Input | This will hold input csv files that the code will use |
| 2\_InterimCode | This will hold our code as we work on it. |
| 3\_Output | This will hold the output files such as generated graphs |
| 4\_DB | Database location |
| 5\_InterimReport | This will be any reports as they are being worked on, including earlier versions |
| 6\_FinalCode | This will be the final code copied from 2\_InterimCode, with the correct filename that states the order to execute |
| 7\_FinalReport | The final report that will be submitted. |

# Database standards

The following is the database standards:

1. All table names will be lower case and separated by underscore
2. All column names will be lower case and separated by underscore, with the exception of the “key word” ID

We need to discuss the final name of the database.

On the following page, I modeled the database as per the information in the proposal. Is this correct? Let’s have a database dictionary document (to be included in our final document) which would be a living document as we document the structure of our database:

Data Model:

