Technical Documentation for NTIA Digital Nation Data Analysis

Introduction:

The Digital Nation Data Analysis project was assigned to a three member team belonging to the University of Washington iSchool. As a part of their coursework, the students took on the challenge of analyzing the Digital Nation Data collected in July of the year 2015.The project involved mining a tremendous dataset describing computer and Internet use in America to tell new stories, inform state and county level policies, and model county-level Internet usage.The team was tasked with visualizing their analysis to facilitate richer dialog between policy analysts and policy makers. The sponsors for this project was NTIA. The outcome of this project is intended to support National Telecommunication and Information Administration (NTIA)’s effort to improve broadband deployment , advance broadband adoption and shape information policies.The project was assigned to the team in January 2018 and completed by June 2018.

The team handed over three deliverables as a part of the project and this technical documentation as an guide to reproduce the work delivered. . These deliverables were designed to expose data points that can accelerate the goal of expanding broadband access and adoption in America while ensuring internet remains an engine for continued innovation and economic growth. Here are the deliverables:

**Deliverable 1**: **State Broadband Profiles**: Create profiles of computer and internet use for each state, presenting the information clearly using compelling visuals and info-graphics. Create the type of charts and materials that might be used to inform policy at the state level.

Work :

<https://drive.google.com/open?id=1v8sYILuGAf5dtgZRTSC059Y82tPIOFHz-A4upuCmUKg>

**Deliverable 2: Complement the existing visualization:** Build visualizations that integrate and showcase multiple variables and features of the data set in a single visualization. This work is to compliment the existing visualizations that represent a single feature in a state map.

Work: <https://nannxiao.github.io/ntia/>

**Deliverable 3: Find a New Story**: NTIA blogs cover “Understanding Veterans’ Computer and Internet Use,” “Why’s Not Online and Why,” and “The State of the Urban/Rural Digital Divide.” What can you see in the data? Find and tell a new story about Computer and Internet use in America.

Work: <https://public.tableau.com/profile/janani.kumar#!/vizhome/DataStories/Story1>

**Make it Repeatable:** Please leave tools, code, and documentation so that we can run these calculations and produce materials after you leave.

The tools used to build these deliverables were: Python, JavaScript, Tableau, Excel.

The team was unable to work on the following deliverable as county level information was not available in the data.

**Model County Level Computer and Internet Use**: While the data is statistically significant at the state level, many people want data at the County level. When people in Grant County WA and Lincoln County WA see the figures for Washington State, they don’t see their community; they see figures overweight for Seattle and King County. Since the trends in computer and Internet use vary by demographics such as income level, education level, and age, please develop a model to localize the data for Counties. If you can model computer and Internet use at the County level, with the appropriate caveats, develop county level summaries with compelling visuals and infographics.

Team:

The team consists of three students in the Masters in Information Management Program at the iSchool, University of Washington Seattle. They are on track to specialize in data science, business intelligence and project management and graduated in summer 2018. The collective goal of the team was to build data driven products and services that will empower citizens of the world. Here are the names and contact details of the students:

Nan Xiao: <https://www.linkedin.com/in/nan-xiao/>

Shiyi Luo:<https://www.linkedin.com/in/shiyi-luo/>

Janani Kumar:<https://www.linkedin.com/in/janani-kumar/>

Data:

The July 2015 techdocs for the dataset Digital Nation Data Analysis is the meta document for this project. The team’s understanding of the data is derived from there. All further readings in this document contain references to the July 2015 techdocs. The

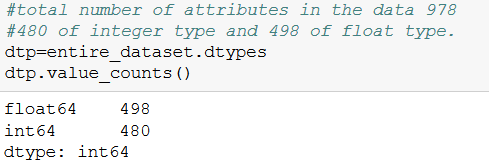
Before delving into the details of how each deliverable was achieved, here is a deep dive on the team’s understanding of the data set.

All scripting and data processing was done using Python in a jupyter notebook using libraries pandas and numpy. You can read more about jupyter notebooks [here](http://jupyter-notebook.readthedocs.io/en/latest/notebook.html).

* The dataset was loaded into a [dataframe](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html). Dataframe allows the dataset to be processed while retaining its table like structure.

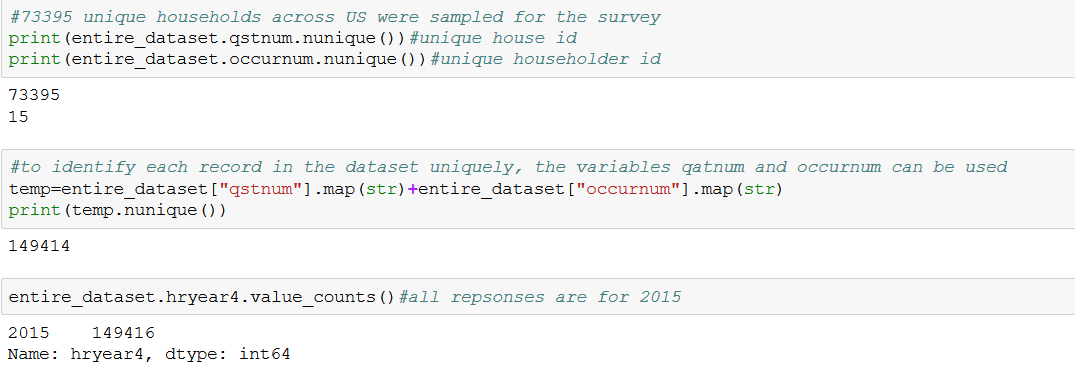


* The dataset has 978 columns. All responses are recorded as numbers. 480 columns are of type integer and 498 columns are of type float.



* 73,395 unique household were interviewed across america. The number of people interviewed per household ranged from 1 to 15. Collectively, the dataset has recorded interviews of 149416 respondents.

To uniquely identify each record in the data set, the columns QSTNUM and OCCURNUM can be used. QSTNUM is used to identify each household uniquely and OCCURNUM can be used identify each person in the household uniquely. Therefore, if the data set id grouped by these two columns, each row can be identified uniquely.



* Each record has different types of weights associated with it. Here are the 4 weights that were explored for this project.

Pwsswgt - Used for most tabulations, controlled to independent estimates for 1) states; 2) origin, sex,and age; and 3) age, race, and sex.

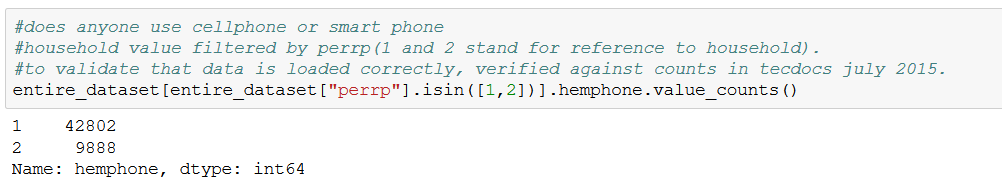
Pwcmpwgt - Family weight: only used for tallying family characteristics

Pwprmwgt - Random Respondent Weight

Hwhhwgt - Household weight used for tallying household characteristics

All household level analysis were computed using the Hwhhwgt weight and person level analysis was computed using Pwsswgt. Random respondent weight and Family weight were not applicable to the computations.

* For household level computations, the response of the reference person of that household was considered as the response of everyone living under that household. These reference persons of the household are usually considered as the head of the household. The column “perrp” is used to identify reference persons in each household.
* To check if the data was loaded correctly, unweighted counts were computed and checked against values presented in the July 2015 technical documentation of the data set.



**How to reproduce Deliverable 1?**

Deliverable 1 was to create profiles of computer and internet use for each state, presenting the information clearly using compelling visuals and info-graphics.

Link to Work : Infographic for the state of Washington <https://drive.google.com/open?id=1v8sYILuGAf5dtgZRTSC059Y82tPIOFHz-A4upuCmUKg>

Data validation and pre-processing steps:

The process includes using appropriate identifier, filtering out the invalid survey responses, and applying necessary weights to different respondents.

To calculate the percentage of people using different internet devices and locations, we used **personal level identifier**. For reasons that people don’t use internet, we used **household level identifier**. Demographic profiles are calculated on **personal identifier** expect for the average household size and annual income which are based on the household identifier. Appropriate **weights** have been applied to different identifier levels.

* How to get data for each state using the Tableau Data Exploration Tool; how to repeat the infographic for other states:

We designed and created the Washington State Computer and Internet User Profile as the sample infographic. The same template can be used to create state level profiles for other states and other updated data source. To reduce the manual process and automize the reproducing work, we also created a Data Exploration Tool using Tableau to help us get the key metric numbers for each state.

*Link to download the Data Exploration Tool:* [*https://drive.google.com/open?id=1X1xRaiks8srFCrNB\_lLZjFfVnfV\_l\_PK*](https://drive.google.com/open?id=1X1xRaiks8srFCrNB_lLZjFfVnfV_l_PK)

The Data Exploration Tool contains the digital nation dataset 2015 and the necessary data aggregation and pre-processing steps. It has two dashboards, the first one contains the computer and internet usage numbers, and the second one contains the demographic profile numbers. Following the **instructions** on the dashboard, one can filter the data by state and get the numbers needed for infographic for any state. To reproduce an infographic for another state, one needs to open the Data Exploration Tool first, filter by that state, get the updated numbers, then follow the instructions to do some simple calculation in order to insert the final numbers into the infographic template.

Please noted that the information covered by the current version of the infographic contains not only our analysis output, and also other metrics provided through the feedbacks on May 18th. Numbers related to the following topics might not be able to re-produced:

1. Locations that children use internet, WA vs. US, school age vs. pre-schooler

2. Digital divides of using internet, WA vs. US

3. Households that are not using internet, WA vs. US

**How to reproduce Deliverable 2?**

Link to Work: <https://nannxiao.github.io/ntia/>

Link to Data Processing Code:

<https://github.com/nannxiao/ntia/blob/gh-pages/Data%20processing-D3_Viz_Code.ipynb>

The data processing code is Python in jupyter notebook format well-commented. The jupyter notebook processed raw data to the customized dataset for our visualizations. The first section contains data processing code for Internet usage map. The function calculates percentage of Internet users by location for specific state, then loops through all states in US. Then we loaded data frame into CSV file. All percentages in first section were weighted using pwsswgt(person level weight parameter).

The second section processed data for pie and bar chart. Because there one more dimension (demographic) added, JSON data format is more suitable than CSV format. Firstly adding two columns for age band and sex in the raw data frame respectively. The function count the number of Internet users of different age band or sex by location. Moreover, the code handles null values (eg. the number of people are over 72 years old use Internet at other location in North Dakota is null in the dataframe) by inserting zero. After getting final data frame, we constructed JSON data structure and loaded data from data frame into JSON. Finally, we copy and paste JSON data to **dashboard2.js** (D3 JavaScript file). All percentages in second section were unweighted because we want to calculate the demographic distribution in people using Internet at different locations, so we count number of people instead of using person level weight parameter.

The data processing code only contains topic of Internet usage by location and demographic information about age band and sex. Future users could add more topics such as Internet usage by device, online activities and more demographic information such as education level (peeduca) and income level(hefaminc) , etc.

**D3 Visualization Code:**

Link: <https://github.com/nannxiao/ntia/tree/gh-pages>

Files in the repository:

* **index.html**: The HTML code for the first section of visualizations: Internet usage by location in US
* **visualization2.js**: D3 JavaScript code for the first section. It is linked with energy.csv file produced by the first section of data processing code above
* **sectiontwo.html**: The HTML code for the second section of visualizations: Internet usage by location and demographic in US dashboard
* **dashboard2.js**: D3 JavaScript code for the second section. It contains JSON data produced by the second section of data processing code above.

**How to reproduce Deliverable 3?**

Link to the exploratory dashboard:<https://public.tableau.com/profile/janani.kumar#!/vizhome/DataStories/Story1>

Link to the data processing document: <https://github.com/nannxiao/ntia/tree/master/Data%20Stories>

The task here was to expose variables that will allow the stakeholders to understand the reasons for digital divide. The following questions were answered from the data set for this purpose.

1. What are the main reasons stated by households across US for not using the Internet?
2. What are the main reasons stated by households for not using the Internet in each state?
3. What privacy and security concerns hold back internet users across US from using the Internet to its full potential?
4. What online activities are hindered due to these Privacy concerns?
5. What are the basis of these fears? That is, how many internet users report an incident of cyber-bullying and cyber crime in each state?

The data processing document has code blocks dedicated to each question that will elaborate on the columns considered for each question, pre-conditions that the respondents had to meet for a valid response and details on weight computation.

The processed data was then loaded into Tableau and published online. The Tableau workbook can be downloaded from [here](https://public.tableau.com/profile/janani.kumar#!/vizhome/DataStories/Story1) for further analysis.

**Next steps:**

When the 2017 data is available, similar data cleaning processes can be carried out. The same questions can be answered using the latest datasets.

**Code repo:**

This is the link to the code repo with all the deliverables and this report.

<https://github.com/nannxiao/ntia>

**Contact details:**

* Github:
  + Janani: @jkumar204
  + Nan: @nannxiao
  + Shiyi: @luos9
* Linkedin:
  + Janani: [https://www.linkedin.com/in/janani-kumar](https://www.linkedin.com/in/janani-kumar/)
  + Nan: [https://www.linkedin.com/in/nan-xiao](https://www.linkedin.com/in/nan-xiao/)
  + Shiyi: https://www.linkedin.com/in/shiyi-luo-66494b93