Mohammad Rahman ([LinkedIn](https://www.linkedin.com/in/mohammadarahman/)) 13 June, 2022

MS, Tech Systems Management (Data Analytics focus)

Stony Brook University

**Independent Data Analysis of the**

**Ethnic and Community Media directory of New York City**

**Preview**

|  |  |
| --- | --- |
| Dataset | Ethnic and Community Media Print, Digital, and Broadcast Outlet Directory |
| Source of Data | NYC Open Data ([click here](https://data.cityofnewyork.us/City-Government/Ethnic-and-Community-Media-Print-Digital-and-Broad/gc4z-q69i)) |
| Provider of Data | Mayor's Office of Operations (OPS) |
| Date Published | 11/4/2020 |
| Date Updated | 03/1/2022 |
| Tool/Technology Used for This Data Analysis | R Programming Language |

**Introduction**

This dataset provides a list of ethnic and community media (ECM) outlets and their descriptions. It has total of 360 observations and 14 variables. Out of these 14 variables, I have selected 6 variables for this analysis purpose and it is because the other variables just include information about outlets’ contact information – which is not significant for this analysis. Here are the 6 variables I am using for this analysis –

|  |  |
| --- | --- |
| **Variables** | **Descriptions** |
| Outlet | The name of the outlet or organization |
| Outlet Format | The format/media in which the outlet is presented |
| Type Of Media | The type/media in which the outlet is presented |
| Languages Served | The languages in which the outlet is published |
| Geographic Focus | The primary geography/geographies targeted by the outlet |
| Communities Served | The primary demographic groups targeted by the outlet |

**Data Cleaning**

First of all, I have worked on cleaning the data as this dataset (the selected variables) has some errors. The major issue with the data is the after/end whitespace in the cells of almost all the variables, especially in the Languages Served variable. Therefore, for example, the English language is written as “English” and “ English” – which is not ideal and needed to be fixed for analysis. Besides, the data collection and entry seem to be inconsistent as well. For instance, the language “Bengali” is written as “Bengali,” “Bangla,” and “Bangladeshi” whereas it should be only either “Bengali” or “Bangla.” Similarly, the Communities Served variable includes “African American” & “African-American” despite they mean the same community. This is also needed to be fixed before analysis. Also, even though very few, there are some wrong data in this dataset. One such example is observation #7 (Akhon Samoy) which is shown as the Thai language served outlet whereas it is actually a Bengali language served outlet. Then, unavailable data are mostly shown as empty cells, however, they are also shown as “-” which is inconsistent.

These are some of the issues that I had to sort out first before doing my analysis.

**Difficulty Of the Dataset and Creation of a New Data-Frame from the Original Dataset**

The Languages Served, Geographic Focus, and Communities Served are the variables that include multiple data within a single cell of the variables. For example, the “Easy 96” outlet (observation #88) served English, Hindi, Bengali, Punjabi, Bangladeshi, Urdu, & Gujarati (a total of 7) languages, and all of the languages are listed by comma (,) in the same cell. This is also done with the Geographic Focus and Communities Served variables for several times. This is certainly an obstacle to do the statistical descriptive analysis of the dataset. To overcome this, I used the str\_split\_fixed function to split up a cell into a fixed number of pieces with either available data or an empty cell. Then I have assigned new variables of all the languages, geography, and communities served. And finally, I have created a new dataset with these three variables to further analyze the data.

Details about this step can be found in the R code.

**Statistical Descriptive Analysis**

This dataset (the variables used in more particular) is all string data types and therefore, I am only doing the analysis to get some factual insights from the dataset. Similarly, due to the limitation/nature of the dataset – no numeric values, I am only using the geom\_bar function of the ggplot2 package to visualize the finding later in the report.

**Outlet**

* There is a total of 360 outlets registered as the ethnic and community media in the NYC

**Outlet Format**

* Here is the breakdown of different outlet format –
* Journal - 4
* Magazine - 33
* Newsletter - 2
* Newspaper - 245
* Radio - 23
* Television – 50
* Other Website - 2
* Website – 1

**Type of Media**

* Here is the breakdown of the type of the media of the outlet
  + Print and Digital - 150
  + Broadcast - 73
  + Digital - 111
  + Print – 26

**Geography Served**

* Unique Geography served by the ECM outlets –11 geography
* Top 5 geography served –
  1. National (81 outlets)
  2. New York City (56)
  3. International (40)
  4. Brooklyn (35)
  5. Queens (34)
* Bottom 3 geography served –

1. Regional (8)
2. Staten Island (2)
3. Far Rockway (1)

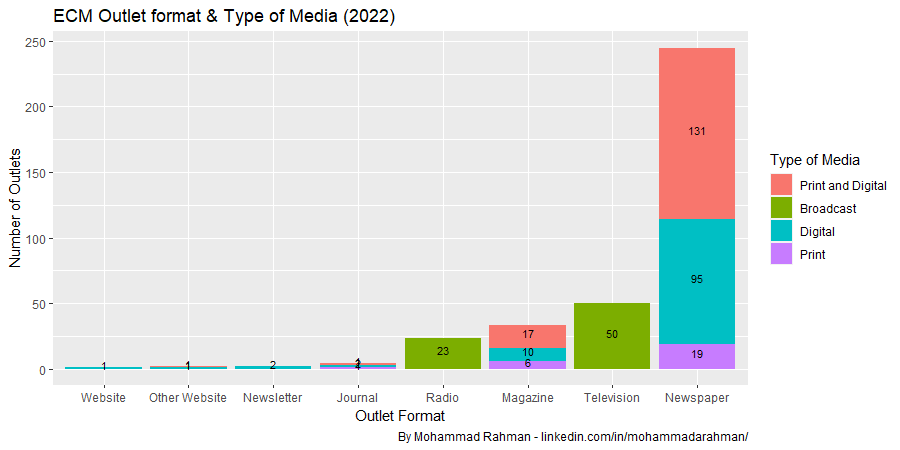
**Languages Served**

* Unique Languages served by the ECM outlets – 35 Languages
* Top 5 geography served –
  1. English (249 outlets)
  2. Bengali & Spanish (22)
  3. Russian (15)
  4. Urdu (14)
  5. Chinese (10)
* Bottom 13 languages are served by just 1 outlet each

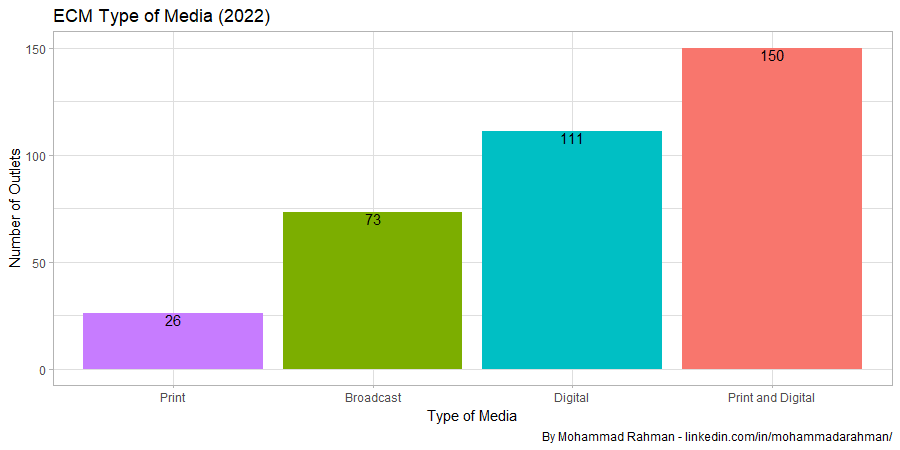
**Communities Served**

* Unique Communities served by the ECM outlets – 75 communities
* Top 5 communities served –
  1. Jewish (43 outlets)
  2. Queens (29)
  3. Brooklyn (23)
  4. Manhattan (16)
  5. African American, Bangladeshi, & Indian (15)
* Bottom 40 communities are served by just 1 outlet each
* 2nd bottom 10 communities are served by 2 outlets each.

**Visualization**

1. **ECM Outlet format & Type of Media**

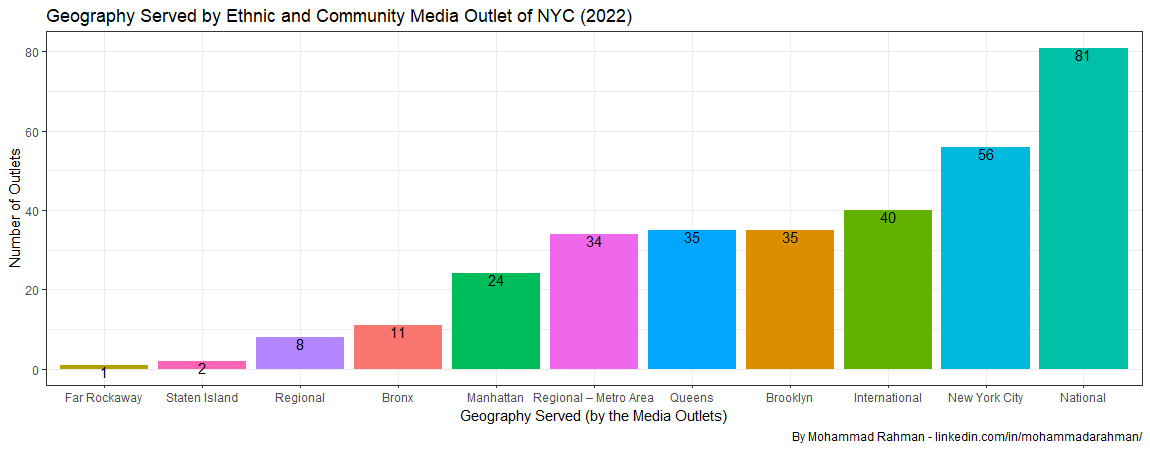
This barograph shows the number of outlets in a different format. Besides, it also shows the type of media that this outlet format is. It is clear that the newspaper has outnumbered the other outlet format with a total of nearly 250 outlets. On the other side, websites, newsletters, and journals are very nominal in number whereas radio, magazine, and television channels are represented by between 25 to 50 outlets.

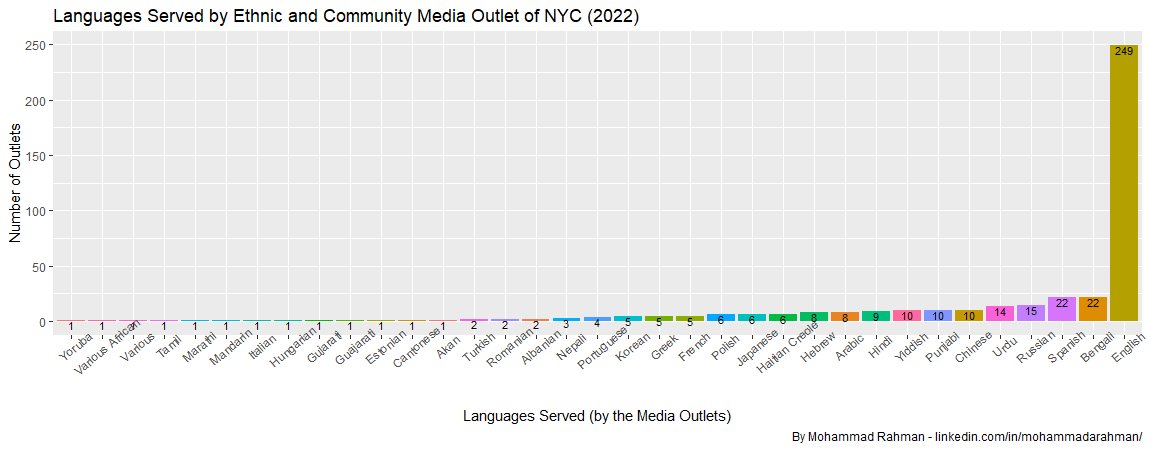
1. **ECM Type of Media**

ECM media outlets are mostly dominated by print and digital publications. The second position is taken by the digital publication whereas the 3rd one is taken by the broadcast media. It can be estimated that the number of print media outlet will decrease, as they are already the lowest type of media circulation, due to the growing availability of smart devices that makes our access to news, entertainment, etc. very efficient and less expensive. A similar analysis in the future can show this trend of which type of media outlet is increasing/decreasing.

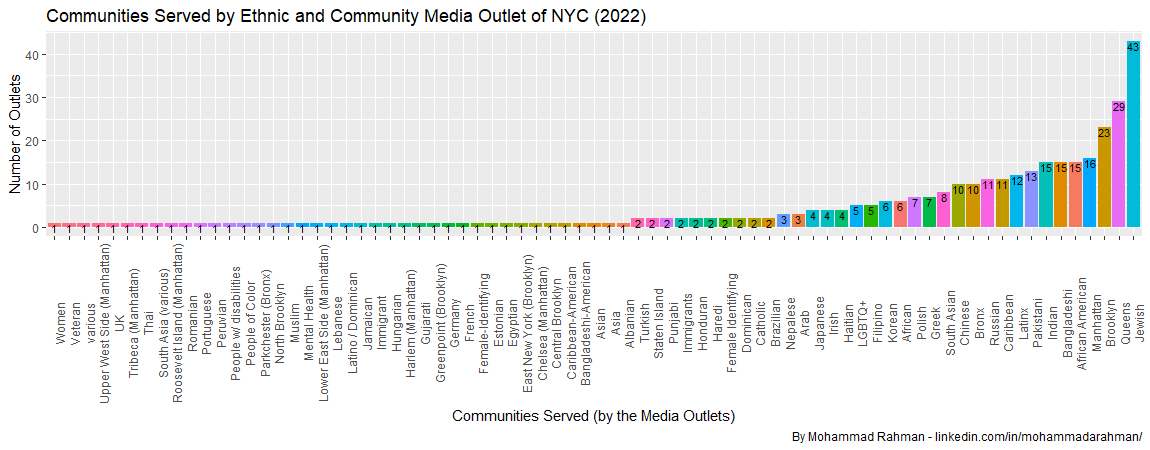
1. **Geography Served by the ECM outlets in NYC**

In the below graph, most of the outlets intended to serve nationally. Then, the entire New York City is served by the 56 outlets. It is interesting to see the 3rd position where ECM media outlets served internationally. In my understanding and analysis of this dataset, this is because there are lots of international media outlets included in the directory. And due to the huge number of immigrants living in New York City, the international media outlets serve them where the outlets are based on the country of origin of these immigrants. Next, the Brooklyn & Queens are served by 35 media outlets each.

1. **Languages Served by the ECM outlets in NYC**

English is unsurprisingly maintaining the graph with 249 outlets publishing their contents in this language. The second position is shared by Spanish & Bengali language with 22 media outlets for each of those two languages. The other languages that follow next in terms of the number of outlets are Russian, Urdu, Chinese, etc. Further analysis can be done to see if the most spoken languages of NYC are represented similarly in the media directory or not.

1. **Communities Served by the ECM outlets in NYC**

****The Jewish community gets the most media coverage from a total of 43 ECM outlets. Then, the second, third, and fourth position is held by Queens, Brooklyn, & Manhattan respectively. The fifth position is shared by African American, Bangladeshi, and Indian communities living in New York City. Like the language served, further analysis can be done to see if the communities are well represented in the ECM directory in accordance with the number of populations of these communities living in New York City.

**Conclusion**

This is an independent preliminary data analysis report to further study and learn about the ethnic and community media outlets as well as the geography, languages, and communities they served. This analysis can be helpful for correcting the existing errors in the dataset as well as being more conscious in data collection and entry since the dataset is being updated gradually contingent upon the inclusion of a new ECM outlet or changes of an existing outlet in the directory. As mentioned earlier, further analysis can be done to compare this dataset with other relevant datasets to study the representation of geography, languages, and communities in New York City.