ANALYTICS IN PRACTICE
DATA PREPARATION

Rajendra.p

RPUSARAP@KENT.EDU

ASSIGNMENT 1

REPORT: -

From this Assignment I learned that Data preparation and Data preprocessing are tough part than applying modeling techniques to data. For, the first two datasets there is a lot of missing values and NA values which are present in few columns. To perform any analytics on this type of data we have to impute the NA values or missing values by knn Imputation or mean or mode. We can fix the dataset by selecting only the important columns in the dataset i,e feature Engineering ,domain knowledge etc. These variables can be identified by having domain knowledge on the datasets and perform some techniques which can identify most important variable in the data and to perform analytics on it. Both us agencies and us companies are connected by a primary key where we can retrieve the data at any point and gain some insights from the data.

I have imported ChicagoTraffic.Json file using Json lite package I have listed out all the columns present in the dataset. There are 23 columns present in the dataset. It took time for me to understand the data and then I find out Total vehicles from 100th to 115th street is 443200 and Traffic at geolocation (41.651861, -87.54501): 4700Traffic at geolocation (41.66836, -87.620176): 8900.

The libraries used to solve this problem are dplyr,tidyverse,jsonlite and rlist libraries.

1. Are there any missing columns?
   No there are no missing columns.
   US agenices

```
{r}
str(us_agencies)# There are no missing columns in agencies dataset
```

```
'data.frame':   1123 obs. of  11 variables:
 $ agency_name     : Factor w/ 70 levels "Administrative Office of
the United States Courts",..: 63 2 3 4 19 19 19 19 19 19 ...
 $ agency_abbrev   : Factor w/ 60 levels "AO","AR","ATL",..: NA 2 3
4 17 17 17 17 17 17 ...
 $ agency_type     : Factor w/ 5 levels "City/County",..: 3 5 1 1 2
2 2 2 2 2 ...
 $ subagency_name  : Factor w/ 56 levels "Agricultural Marketing
Service",..: 26 26 26 26 26 26 7 7 7 7 ...
 $ subagency_abbrev: Factor w/ 53 levels "ACE","AMS","BEA",..: NA
NA NA NA NA 8 8 8 8 ...
 $ url             : Factor w/ 111 levels
"http://catalog.data.gov/dataset?q=organization:((ecab-dol-gov)+OR+
(whd-dol-gov)+OR+(esba-dol-gov)+OR+(ojc-dol-g"| __truncated__,..:
NA 6 91 79 1 1 12 12 12 12 ...
 $ used_by         : Factor w/ 526 levels "(Leg)Cyte","3 Round
Stones, Inc.",..: 89 350 520 101 58 281 180 180 240 240 ...
 $ used_by_category: Factor w/ 20 levels "Aerospace and
Defense",..: 11 8 10 11 8 3 16 16 8 8 ...
 $ used_by_fte     : Factor w/ 8 levels "1-10","1,001-5,000",..: 1
1 1 8 8 4 1 1 1 1 ...
 $ dataset_name    : Factor w/ 506 levels "18 State ALL Payer
Hospital Dataset",..: 421 322 87 NA 354 416 373 238 85 324 ...
 $ dataset_url     : Factor w/ 276 levels
"asterweb.jpl.nasa.gov",..: 38 NA NA NA 131 100 99 99 94 96 ...
```

   US Companies

```r
str(us_companies)
```

```
'data.frame':   529 obs. of  22 variables:
 $ company_name_id    : Factor w/ 529 levels "3-round-stones-
inc",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ company_name       : Factor w/ 529 levels "(Leg)Cyte","3 Round
Stones, Inc.",..: 2 3 4 5 6 7 8 9 10 11 ...
 $ url                : Factor w/ 528 levels "abtassoc.com",..: 29
387 457 1 112 113 114 31 115 116 ...
 $ year_founded       : int  2010 2014 2007 1965 1999 1989 1962
1969 2001 2009 ...
 $ city               : Factor w/ 202 levels " Philadelphia",..:
189 1 66 32 156 38 173 89 80 153 ...
 $ state              : Factor w/ 39 levels "AL","AR","AZ",..: 7 31
36 15 4 11 31 2 4 4 ...
 $ country            : Factor w/ 1 level "us": 1 1 1 1 1 1 1 1 1 1
...
 $ zip_code           : int  20004 19087 22003 2138 94583 60601
16803 72201 92618 95510 ...
 $ full_time_employees: Factor w/ 8 levels "1-10","1,001-5,000",..:
1 8 1 2 7 3 5 6 4 3 ...
 $ company_type       : Factor w/ 9 levels "Nonprofit","nonprofit +
commercial spinoff",..: 6 6 6 6 6 7 6 7 6 7 ...
 $ company_category   : Factor w/ 20 levels "Aerospace and
Defense",..: 3 8 3 18 11 NA 7 3 2 3 ...
 $ revenue_source     : Factor w/ 100 levels "Advertising",..: 37
53 97 26 71 71 87 71 71 71 ...
 $ business_model     : Factor w/ 27 levels "academia","Business to
Business",..: 4 2 8 NA 17 2 8 2 2 21 ...
 $ social_impact      : Factor w/ 12 levels "Citizen engagement and
```

2. Are there any missing column names or errors in the column names? If so, name those columns.
A. No missing column name errors

```r
colnames(us_agencies)
colnames(us_companies)
```

```
 [1] "agency_name"       "agency_abbrev"     "agency_type"
 [4] "subagency_name"    "subagency_abbrev" "url"
 [7] "used_by"           "used_by_category" "used_by_fte"
[10] "dataset_name"      "dataset_url"
 [1] "company_name_id"      "company_name"
 [3] "url"                  "year_founded"
 [5] "city"                 "state"
 [7] "country"              "zip_code"
 [9] "full_time_employees" "company_type"
[11] "company_category"     "revenue_source"
[13] "business_model"       "social_impact"
[15] "description"          "description_short"
[17] "source_count"         "data_types"
[19] "example_uses"         "data_impacts"
[21] "financial_info"       "last_updated"
```

3. Are there any values in the columns missing?
Missing columns in us agencies

```
df <- data.frame(us_agencies)
sum(is.na(df))
colSums(is.na(df))
```

```
[1] 2718
      agency_name     agency_abbrev       agency_type
                0               313                 0
   subagency_name  subagency_abbrev                url
                0               709               292
          used_by  used_by_category       used_by_fte
                0                 4                44
     dataset_name       dataset_url
              548               808
```

Missing columns in us companies

```
df1 <- data.frame(us_companies)
sum(is.na(df1))
colSums(is.na(df1))
```

```
[1] 2315
    company_name_id         company_name                  url
                  0                    0                    0
      year_founded                 city                state
                  1                   33                    0
            country             zip_code full_time_employees
                  0                   37                   29
       company_type     company_category      revenue_source
                 16                    3                   10
     business_model       social_impact          description
                 76                  512                    0
  description_short         source_count           data_types
                  0                  303                  387
       example_uses         data_impacts       financial_info
                521                    0                  387
       last_updated
                  0
```

4 How is data organized in each column? Is it properly organized?

NO, Data is not properly organized in both the data sets.
US COMPANIES

```
summary(us_companies)
```

```
          company_name_id                company_name
 3-round-stones-inc:  1    (Leg)Cyte           :  1
 48-factoring-inc  :  1    3 Round Stones, Inc.:  1
 5psolutions       :  1    48 Factoring Inc.   :  1
 abt-associates    :  1    5PSolutions         :  1
 accela            :  1    Abt Associates      :  1
 accenture         :  1    Accela              :  1
 (Other)           :523    (Other)             :523
              url         year_founded              city
 www.careset.com :  2   Min.   :1799   New York      : 83
 abtassoc.com    :  1   1st Qu.:1994   San Francisco: 45
 apextechllc.com :  1   Median :2007   Boston        : 21
 appallicious.com:  1   Mean   :1993   Chicago       : 17
 asset4.com      :  1   3rd Qu.:2010   Washington    : 17
 auntbertha.com  :  1   Max.   :2015   (Other)       :313
 (Other)         :522   NA's   :1      NA's          : 33
     state      country      zip_code       full_time_employees
 CA     :132   us:529   Min.   :    0   1-10       :143
 NY     :106            1st Qu.:10019   11-50      :115
 MA     : 42            Median :37026   51-200     : 93
 IL     : 26            Mean   :47456   10,001+    : 56
 DC     : 25            3rd Qu.:94025   1,001-5,000: 30
 WA     : 25            Max.   :98144   (Other)    : 63
 (Other):173            NA's   :37      NA's       : 29
      company_type                    company_category
 Private    :396   Data/Technology           : 97
 Public     : 92   Finance & Investment      : 75
 Nonprofit  : 15   Business & Legal Services: 44
 partnership:  4   Governance                : 43
 Partnership:  2   Healthcare                : 40
 (Other)    :  4   (Other)                   :227
 NA's       : 16   NA's                      :  3
                                  revenue_source
```

US AGENICES

```
summary(us_agencies)
```

```
                                        agency_name  agency_abbrev
 Department of Commerce                       :194   USDC    :194
 Multiple government open data sources        :154   HHS     :147
 Department of Health and Human Services      :147   EPA     : 48
 Multiple city and local data sources         : 74   SEC     : 44
 U.S. Environmental Protection Agency         : 48   DOE     : 41
 Securities and Exchange Commission           : 44   (Other):336
 (Other)                                      :462   NA's    :313
            agency_type
 City/County       :135
 Federal           :760
 Federal Open Data:  1
 Other             :154
 State             : 73


                                              subagency_name
 General                                                :707
 US Census Bureau                                       : 83
 National Oceanic and Atmospheric Administration        : 53
 National Institutes of Health                          : 37
 Centers for Medicare and Medicaid Services             : 27
 Bureau of Labor Statistics                             : 25
 (Other)                                                :191
 subagency_abbrev
 Census  : 83
 NOAA    : 53
 NIH     : 37
 CMS     : 27
 BLS     : 25
 (Other):189
 NA's    :709
                                                        url
```

5) Is data in the proper shape for further analysis? If not, why? Explain.

A) Data is not in good shape for further analysis.In dataset like us_agencies columns like dataset_url has many missing values.used_by_fte data is not organized.In datasetlike us_companies colums likefull_time_ employees,source_count,data_impacts has unreadable characters and is not properly parsed.financial_info has lot of missing values almost 73%.

6) How will you fix this dataset? Describe the methods you will use to fix this dataset for further analysis? It can be missing values, NAs, etc.

A) We can fix the dataset by selecting the important columns for example to deal with NA values we can use the technique like knn imputation and other techniques to impute the NA and missing values with most appropriate values.

```
# New clean data frame by droping unwanted variables without omiting
NA Values
agencies_new<-us_agencies %>%
  select("agency_name","agency_type","subagency_name","used_by",
"used_by_category", "dataset_name") %>%
  rename(user = used_by,
         user_category = used_by_category,
         dataset = dataset_name)
str(agencies_new)
```

```
'data.frame':   1123 obs. of  6 variables:
 $ agency_name   : Factor w/ 70 levels "Administrative Office of
the United States Courts",..: 63 2 3 4 19 19 19 19 19 19 ...
 $ agency_type   : Factor w/ 5 levels "City/County",..: 3 5 1 1 2 2
2 2 2 2 ...
 $ subagency_name: Factor w/ 56 levels "Agricultural Marketing
Service",..: 26 26 26 26 26 26 7 7 7 7 ...
 $ user          : Factor w/ 526 levels "(Leg)Cyte","3 Round
Stones, Inc.",..: 89 350 520 101 58 281 180 180 240 240 ...
 $ user_category : Factor w/ 20 levels "Aerospace and Defense",..:
11 8 10 11 8 3 16 16 8 8 ...
 $ dataset       : Factor w/ 506 levels "18 State ALL Payer
Hospital Dataset",..: 421 322 87 NA 354 416 373 238 85 324 ...
```

```{r}
#Filtering dataset to remove columns
companies_clean<-subset(us_companies, select =
-c(1,7,9,14,15,17:21))%>%
                  rename(us_state = state)

head(companies_clean)
```

	company_name <fctr>
1	3 Round Stones, Inc.
2	48 Factoring Inc.
3	5PSolutions
4	Abt Associates
5	Accela
6	Accenture

6 rows | 1-2 of 12 columns

7) How are the two datasets linked to each other? Is there a common “primary key” to connect the two datasets?

A) Yes, there is a primary key that can connect both the datasets. We have renamed used by in agencies to company name in the agencies data set. This company name variable in the agencies

data set is the primary key to the company's data set. We can connect two data sets using the company name data set.

EXERCISE 2

1. How many variables are in the dataset?

A) There are 23 variables in the dataset.

```
{r}
#Importing JSON dataset file
traffic<-fromJSON("ChicagoTraffic.json")

#Counting and Naming all the columns
col_names<-(list.stack(
  list.select(traffic,traffic$meta$view$columns$name)))
print(unique(col_names))
#We have 23 columns in this dataset
```

	V1 <chr>
1	sid
2	id
3	position
4	created_at
5	created_meta
6	updated_at
7	updated_meta
8	meta
9	ID
10	Traffic Volume Count Location Address

1-10 of 23 rows Previous 1 2 3 Next

2. Name all the variables?

The names of 23 variables are: sid, id, position, created_at, created_meta, updated-at, updated_meta, meta, Id, traffic volume count Location Address, Total Passing Vehicle Volume, Vehicle Volume By Each Direction of Traffic, Latitude, Longitude, Location, ZBoundaries-zip codes, Community Areas, Zip Codes, Census Tracks, Wards, @Computed_region.

	V1 <chr>
20	Zip Codes
21	Census Tracts
22	Wards
23	:@computed_region_awaf_s7ux

	V1 <chr>
1	sid
2	id
3	position
4	created_at
5	created_meta
6	updated_at
7	updated_meta
8	meta
9	ID
10	Traffic Volume Count Location Address
11	Street
12	Date of Count
13	Total Passing Vehicle Volume
14	Vehicle Volume By Each Direction of Traffic
15	Latitude
16	Longitude
17	Location
18	Boundaries - ZIP Codes
19	Community Areas

3. What is the total traffic of vehicles on 100th street to 115th street?

Total vehicles from 100th to 115th street is 443200

```r
#To find traffic count from 100th to 115th Street
x<-list("100th St","101 St","102 St","103rd St","104th St","105th
St","106th St","107th St","108th St","109th St","110th St","111th
St","112th St","113th St","114th St","115th St")
total_traffic<-0
for (j in 1:16){
  for(i in 1:1279){
    if(traffic$data[[i]][[11]] == x[j])
    {
      total_traffic <- total_traffic +
as.numeric(traffic$data[[i]][[13]])
    }
  }
}


print(paste("Total traffic from 100th street to 115th
street:",total_traffic))
```

```
[1] "Total traffic from 100th street to 115th street: 443200"
```

4. What is the total traffic of vehicles on geolocations, (41.651861, -87.54501) and (41.66836, -87.620176)

Traffic at geolocation (41.651861, -87.54501): 4700

Traffic at geolocation (41.66836, -87.620176): 8900

```
  {r}
#To find traffic count at specified Geolocation
for(i in 1:1279){
  if(traffic$data[[i]][[15]]=="41.651861" &
traffic$data[[i]][[16]]=="-87.54501")
  {
    print(paste("Traffic at geolocation (41.651861,
-87.54501):",traffic$data[[i]][[13]]))
    print(i)
  }
  else if(traffic$data[[i]][[15]]=="41.66836" &
traffic$data[[i]][[16]]=="-87.620176")
  {
    print(paste("Traffic at geolocation (41.66836,
-87.620176):",traffic$data[[i]][[13]]))
    print(paste(i))
  }
}
```

```
[1] "Traffic at geolocation (41.651861, -87.54501): 4700"
[1] 149
[1] "Traffic at geolocation (41.66836, -87.620176): 8900"
[1] "965"
```