Use Case

- *To identify annual package/salary for R&D company employees for sectors such as Private Sector, government Sector, Higher Education Sector, and Public Research Institutes for roles including Researchers, Technicians, and Other Supporting Staff from year 2011 to 2014.*

The public data sets used are 1 - "Research and Development Manpower Headcount by Sector" and 2 - "Research and Development Expenditure by Type of Cost", accessed from *data.gov.sg*.

The first data set provides information on total headcount of R&D manpower. Data entry documenting headcount for PhD, Masters, Bachelors, Technicians, and Other Supporting Staff are of interest for this analysis. Therefore, the data set is manipulated to present this information. The assumption here is that Researchers include only PhD, Masters and Bachelors. Therefore, the data set is transformed to be compatible with the second data set.

Second data set is for R&D expenditure. Manpower expenditure is the data entry needs to be extracted. Therefore, EOM (Expenditure of Manpower) for Researchers, Technicians, and Other Supporting Staff are selected.

A dimensionality check on the data is performed to truncate incomplete or unnecessary data.

The first data set elements are then transformed by mathematically dividing the second data set elements to obtain the estimated average annual package per Researcher, Technician, or Other Supporting Staff. The results retrieved characterise the annual package such as the annual salary and benefits the staff receive according to different sectors.

From an employee's perspective, this could be a benchmark or guidance on which sectors to work if he/she is interested in the R&D field, for a better monetary reward for various roles (Researcher, Technician, or Other Supporting Staff). For corporations, this provides an insight to the market trend of R&D investment by the government and industries on workforce development. Therefore, aiding corporate level decision making on whether to synergise with government backed R&D centres for its own R&D activities.

Database Schema

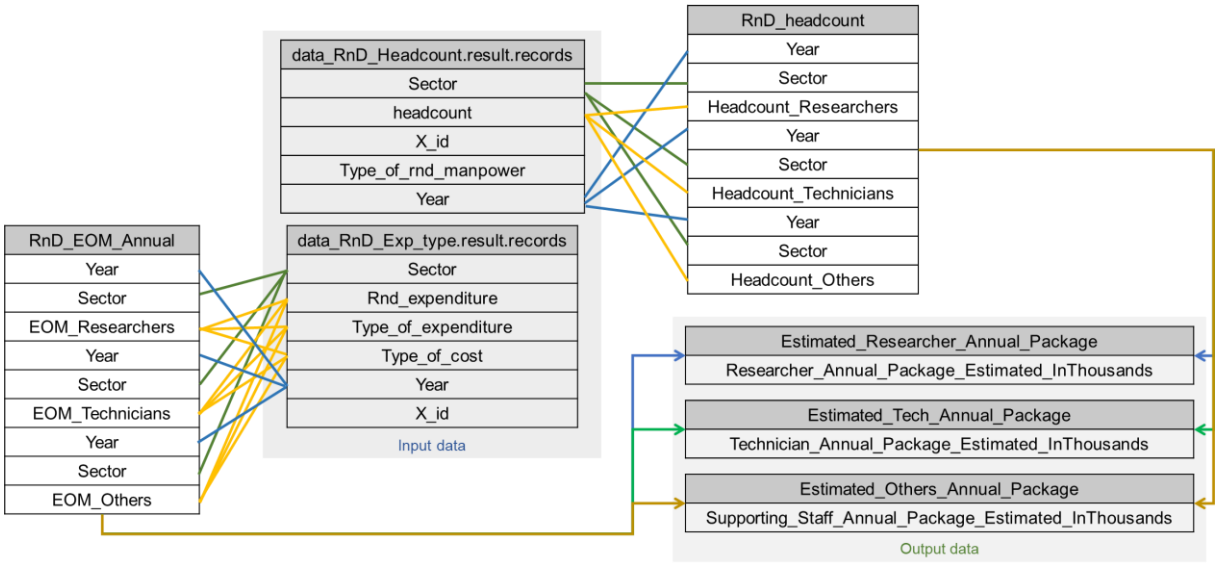

Figure 1. Input output data schema

Appendices

- The script is developed in Matlab using mostly C language.
- APIs such as RESTful is used for HTTP methods. None of the toolboxes is used. Standard functions such as strcmp, find, cellfun, str2double, cell2table, join, writetable, flip, bar, label and legends are used throughout the codes.
- The online data set resource weblinks are provided below
    - "Research and Development Manpower Headcount by Sector" (100 readable data entries) https://data.gov.sg/api/action/datastore_search?resource_id=a6fb9294-f0d2-4851-8b18-1e90ce54f130
    - "Research and Development Expenditure by Type of Cost" (100 readable data entries) https://data.gov.sg/api/action/datastore_search?resource_id=2778094d-0804-48a3-957e-31ab777eb306
- The generated graphs by the script are provided below for reference
    - For Researcher

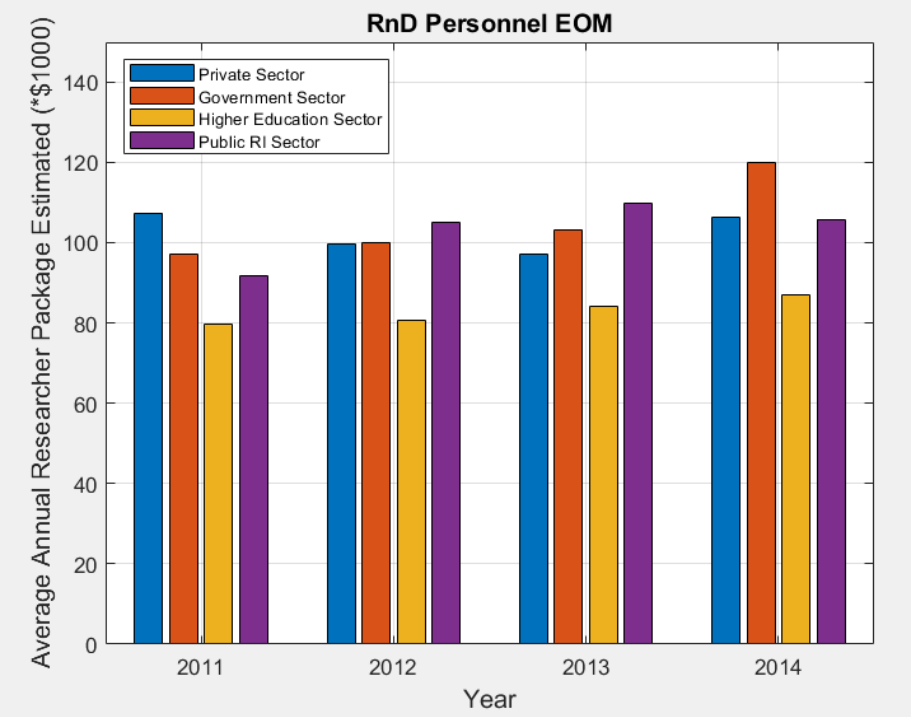

Figure 2. Annual package information for Researcher

    - For Technician

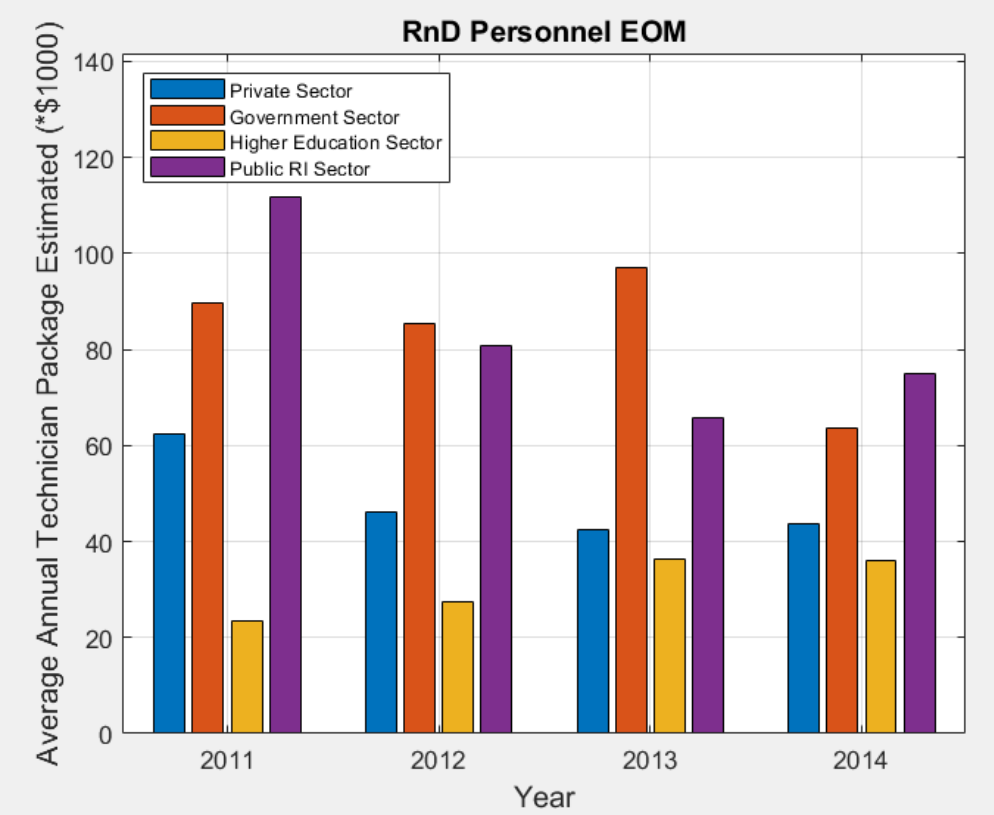

Figure 3. Annual package information for Technician

- For Other Supporting Staff

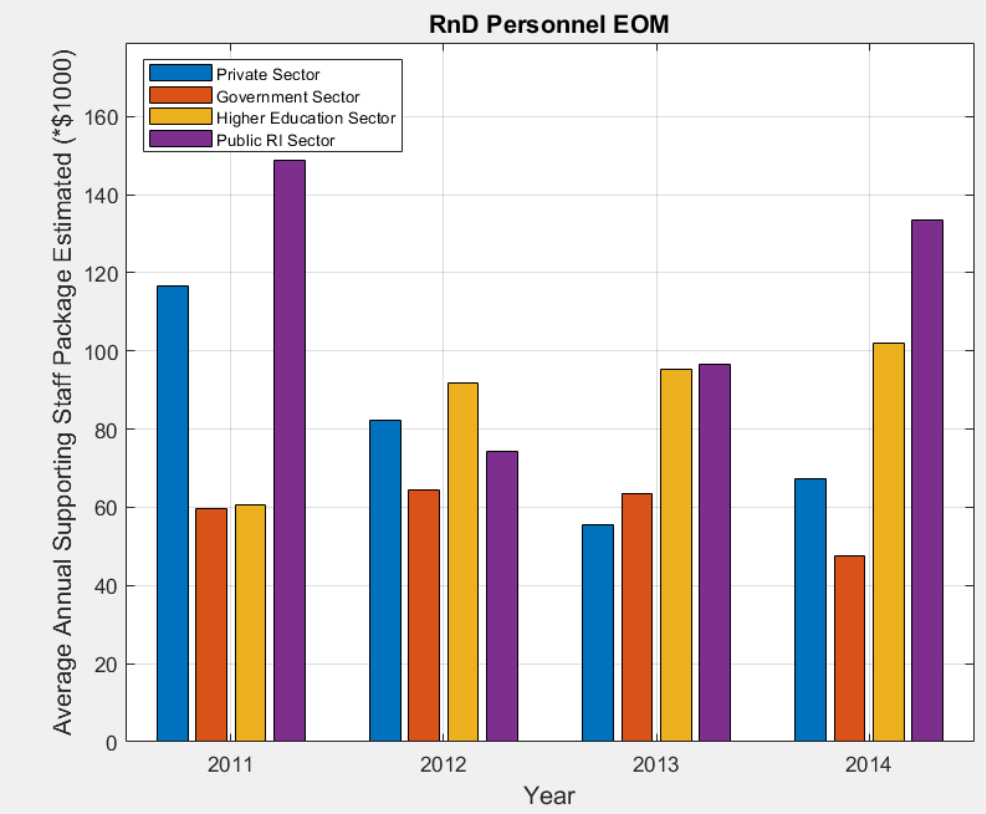

Figure 4. Annual package information for Other Supporting Staff

- Source codes

```
clc; clear; close all

%% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  ReadMe  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% %%
% - This is a script developed to read public data sets from weblinks(compatible for
source data updates
%   for data analytics, as instructed by the Technical Assessment 1 IMDA
% - The script runs on Matlab (Only tested with 2019b version).
% - It reads data links shown in the codes and automatically processes the
%   data which then produces the graphs illustrating the analysis.
% - The public data sets used are "Research and Development Manpower Headcount by Sector"
%   -> https://data.gov.sg/dataset/research-and-development-manpower-headcount-by-sector-
2014
%   & "Research and Development Expenditure by Type of Cost".
%   -> https://data.gov.sg/dataset/research-and-development-expenditure-by-type-of-cost
% - The outpt file and graphs depict the estimated average annual salary
%   researchers, technicians, support staff receive from 2011 to 2014 based
%   on different sectors.
% - Author: Liu JiaYu <jiayuworks(æt)gmail.com>
%   Last modified on 9 Oct 2019 3:19 PM

%% Data Download %%
data_RnD_Headcount =
webread('https://data.gov.sg/api/action/datastore_search?resource_id=a6fb9294-f0d2-4851-
8b18-1e90ce54f130');
data_RnD_Exp_type =
webread('https://data.gov.sg/api/action/datastore_search?resource_id=2778094d-0804-48a3-
957e-31ab777eb306');

%%%% Data Manipulation & Transformation %%%%
```

```
tic

%% R&D Headcount Database %%

%%% Initialisation %%%
column_space = 3;   %% 3 columns of info needed from source data. Can be changed based on
scenarios
annual_researchers_by_sector = cell(data_RnD_Headcount.result.total,column_space);
annual_technicians_by_sector = cell(data_RnD_Headcount.result.total,column_space);
annual_others_by_sector = cell(data_RnD_Headcount.result.total,column_space);
num_researchers = 0;
num_technicians = 0;
num_others = 0;
i=1;
rows_of_data = 100;     %% manually adjusted based on # of data seen

    for j=1:(rows_of_data)
            %%% check for type of manpower based on sector %%%
          check_researchers = sum
(strcmp(data_RnD_Headcount.result.records(j).type_of_rnd_manpower,"PhD") +...

strcmp(data_RnD_Headcount.result.records(j).type_of_rnd_manpower,"Masters")+...

strcmp(data_RnD_Headcount.result.records(j).type_of_rnd_manpower,"Bachelors"));
          check_technicians =
strcmp(data_RnD_Headcount.result.records(j).type_of_rnd_manpower,"Technicians");
          check_others =
strcmp(data_RnD_Headcount.result.records(j).type_of_rnd_manpower,"Other Supporting
Staff");
```

```
            %%% check for sector and year with the neighbouring data row (taking
consideration of last row) %%%
            %%% sector is the slicing criterion %%%
         if j~=rows_of_data
         check_sector= strcmp(data_RnD_Headcount.result.records(j).sector,
data_RnD_Headcount.result.records(j+1).sector);
         else
          check_sector= strcmp(data_RnD_Headcount.result.records(j).sector, "There is no
more data");
         end
            if check_researchers
            num_researchers = str2double(data_RnD_Headcount.result.records(j).headcount)
+ num_researchers;
            elseif check_technicians
            num_technicians = str2double(data_RnD_Headcount.result.records(j).headcount)
+ num_technicians;     %% expandable to more technician categories.
            elseif check_others
            num_others = str2double(data_RnD_Headcount.result.records(j).headcount) +
num_others;   %% expandable to more categories
            end

        %%% Data reformat %%%
            if ~check_sector
            annual_researchers_by_sector(i,:) =
[str2double(data_RnD_Headcount.result.records(j).year)
cellstr(data_RnD_Headcount.result.records(j).sector) num_researchers];
            annual_technicians_by_sector(i,:) =
[str2double(data_RnD_Headcount.result.records(j).year)
cellstr(data_RnD_Headcount.result.records(j).sector) num_technicians];
```

```
            annual_others_by_sector(i,:) =
[str2double(data_RnD_Headcount.result.records(j).year)
cellstr(data_RnD_Headcount.result.records(j).sector) num_others];
            i = i+1;

            num_researchers = 0;
            num_technicians = 0;
            num_others = 0;
            end
    end
       %%% tidying up data - removeing null cells %%%
       zerodata_index_rh = find(cellfun('isempty',annual_researchers_by_sector));
       annual_researchers_by_sector = annual_researchers_by_sector(1:(zerodata_index_rh-
1),1:column_space);
       zerodata_index_th = find(cellfun('isempty',annual_technicians_by_sector));
       annual_technicians_by_sector = annual_technicians_by_sector(1:(zerodata_index_th-
1),1:column_space);
       zerodata_index_oh = find(cellfun('isempty',annual_others_by_sector));
       annual_others_by_sector = annual_others_by_sector(1:(zerodata_index_oh-
1),1:column_space);

%% R&D Expenditure_by_Type Database %%

%%% Initialisation %%%
researchers_EOM_by_sector_annual = cell(data_RnD_Exp_type.result.total,column_space);
technician_EOM_by_sector_annual = cell(data_RnD_Exp_type.result.total,column_space);
others_EOM_by_sector_annual = cell(data_RnD_Exp_type.result.total,column_space);
k=1;
rows_of_data = 100;     %% manually adjusted based on # of data seen
```

```
    for j=1:(rows_of_data)
        %%% check for all data row and extract those concerning EOM %%%
          check_researchers =
strcmp(data_RnD_Exp_type.result.records(j).type_of_cost,"Researchers");
          check_technicians =
strcmp(data_RnD_Exp_type.result.records(j).type_of_cost,"Technicians");
          check_others = strcmp(data_RnD_Exp_type.result.records(j).type_of_cost,"Other
Supporting Staff");
          check_exp_EOM = strcmp(data_RnD_Exp_type.result.records(j).type_of_expenditure,
"Manpower Expenditure");

        %%% Data reformat %%%
            if check_exp_EOM && check_researchers
            researchers_EOM_by_sector_annual(k,:) =
[str2double(data_RnD_Exp_type.result.records(j).year)
cellstr(data_RnD_Exp_type.result.records(j).sector)
str2double(data_RnD_Exp_type.result.records(j).rnd_expenditure)];
            elseif check_exp_EOM && check_technicians
            technician_EOM_by_sector_annual(k,:) =
[str2double(data_RnD_Exp_type.result.records(j).year)
cellstr(data_RnD_Exp_type.result.records(j).sector)
str2double(data_RnD_Exp_type.result.records(j).rnd_expenditure)];
            elseif check_exp_EOM && check_others
            others_EOM_by_sector_annual(k,:) =
[str2double(data_RnD_Exp_type.result.records(j).year)
cellstr(data_RnD_Exp_type.result.records(j).sector)
str2double(data_RnD_Exp_type.result.records(j).rnd_expenditure)];
            k = k+1;
            end
    end
```

```
        %%% tidying up data - removeing null cells %%%
        zerodata_index_r = find(cellfun('isempty',researchers_EOM_by_sector_annual));
        researchers_EOM_by_sector_annual =
researchers_EOM_by_sector_annual(1:(zerodata_index_r-1),1:column_space);
        zerodata_index_t = find(cellfun('isempty',technician_EOM_by_sector_annual));
        technician_EOM_by_sector_annual =
technician_EOM_by_sector_annual(1:(zerodata_index_t-1),1:column_space);
        zerodata_index_o = find(cellfun('isempty',others_EOM_by_sector_annual));
        others_EOM_by_sector_annual = others_EOM_by_sector_annual(1:(zerodata_index_o-
1),1:column_space);

%% R&D Personnel Annual Overall Expenses %%

        %%% organising data & truncating incomplete/unnecessary data %%%
        dim_min_hc = min([(size(annual_researchers_by_sector));
(size(annual_technicians_by_sector)); (size(annual_others_by_sector))]);
        dim_min_exp = min([(size(researchers_EOM_by_sector_annual));
(size(technician_EOM_by_sector_annual)); (size(others_EOM_by_sector_annual))]);
        dim_min = min(dim_min_hc, dim_min_exp);
        Check_Dim = logical(dim_min_hc==dim_min_exp);
       if Check_Dim(1,1)==1 && Check_Dim(1,2)==1    %IF they are of the same dimensions

        %%% Combined processed data into a table %%%
        RnD_headcount = [annual_researchers_by_sector(1:(dim_min(1,1)),1:column_space)
annual_technicians_by_sector(1:(dim_min(1,1)),1:column_space)
annual_others_by_sector(1:(dim_min(1,1)),1:column_space)];
        RnD_headcount = cell2table(RnD_headcount,'VariableNames',{'Year' 'Sector'
'Headcount_Researchers'...
            '_Year_' '_Sector_' 'Headcount_Technicians' '__Year__' '__Sector__'
'Headcount_Others'});
```

```
        RnD_EOM_Annual =
[researchers_EOM_by_sector_annual(1:(dim_min(1,1)),1:column_space)
technician_EOM_by_sector_annual(1:(dim_min(1,1)),1:column_space)
others_EOM_by_sector_annual(1:(dim_min(1,1)),1:column_space)];
        RnD_EOM_Annual = cell2table(RnD_EOM_Annual,'VariableNames',{'Year' 'Sector'
'EOM_Researchers'...
            '_Year_' '_Sector_' 'EOM_Technicians' '__Year__' '__Sector__' 'EOM_Others'});

        RnD_Personnel_EOM = join (RnD_headcount, RnD_EOM_Annual);

        %%% Annual overall expenses for Researcher %%%
        Estimated_Researcher_Annual_Package =
1000*(RnD_Personnel_EOM.EOM_Researchers./RnD_Personnel_EOM.Headcount_Researchers);
        Estimated_Researcher_Annual_Package=
round(Estimated_Researcher_Annual_Package,3);
        Estimated_Researcher_Annual_Package =
array2table(Estimated_Researcher_Annual_Package,'VariableNames',{'Researcher_Annual_Packa
ge_Estimated_InThousands'});

        Estimated_Tech_Annual_Package =
1000*(RnD_Personnel_EOM.EOM_Technicians./RnD_Personnel_EOM.Headcount_Technicians);
        Estimated_Tech_Annual_Package= round(Estimated_Tech_Annual_Package,3);
        Estimated_Tech_Annual_Package =
array2table(Estimated_Tech_Annual_Package,'VariableNames',{'Technician_Annual_Package_Est
imated_InThousands'});

        Estimated_Others_Annual_Package =
1000*(RnD_Personnel_EOM.EOM_Others./RnD_Personnel_EOM.Headcount_Others);
        Estimated_Others_Annual_Package= round(Estimated_Others_Annual_Package,3);
```

```
        Estimated_Others_Annual_Package =
array2table(Estimated_Others_Annual_Package,'VariableNames',{'Supporting_Staff_Annual_Pac
kage_Estimated_InThousands'});
        %%% Append to main table %%%
        RnD_Personnel_EOM = [RnD_Personnel_EOM, Estimated_Researcher_Annual_Package,
Estimated_Tech_Annual_Package, Estimated_Others_Annual_Package];
       else
            error ('The dimensions of the data do not match. Unable to process further.
\nPlease relook at the source datasets');
       end

%% Output File -> {RnD_Personnel_EOM.xlsx} %%
writetable(RnD_Personnel_EOM,'RnD_Personnel_EOM.xlsx');

%% Graphs Generation %%
data(:,1,1)=RnD_Personnel_EOM.Researcher_Annual_Package_Estimated_InThousands;
data(:,1,2)=RnD_Personnel_EOM.Technician_Annual_Package_Estimated_InThousands;
data(:,1,3)=RnD_Personnel_EOM.Supporting_Staff_Annual_Package_Estimated_InThousands;
for v=1:3 %%% loops for all 3 staff categories
p=1;q=1;w=1;e=1;
     for m = 1:dim_min(1,1)
         if strcmp(RnD_Personnel_EOM.Sector(m),"Private Sector")
             hist_PS(p,1) = data(m,1,v);
             p=p+1;
         elseif strcmp(RnD_Personnel_EOM.Sector(m),"Government Sector")
             hist_GS(q,1) = data(m,1,v);
             q=q+1;
         elseif strcmp(RnD_Personnel_EOM.Sector(m),"Higher Education Sector")
             hist_HES(e,1) = data(m,1,v);
             e=e+1;
```

```
        elseif strcmp(RnD_Personnel_EOM.Sector(m),"Public Research Institutes")
            hist_PRI(w,1) = data(m,1,v);
            w=w+1;
        end
    end
    hist_PS = flip(hist_PS)';
    hist_GS = flip(hist_GS)';
    hist_HES = flip(hist_HES)';
    hist_PRI = flip(hist_PRI)';
    cnt=1;
    year2011 = [hist_PS(1,cnt) hist_GS(1,cnt) hist_HES(1,cnt) hist_PRI(1,cnt)];
    cnt=2;
    year2012 = [hist_PS(1,cnt) hist_GS(1,cnt) hist_HES(1,cnt) hist_PRI(1,cnt)];
    cnt=3;
    year2013 = [hist_PS(1,cnt) hist_GS(1,cnt) hist_HES(1,cnt) hist_PRI(1,cnt)];
    cnt=4;
    year2014 = [hist_PS(1,cnt) hist_GS(1,cnt) hist_HES(1,cnt) hist_PRI(1,cnt)];

    figure()
    years = 2011:1:2014;
    AP = [year2011;year2012;year2013;year2014];
    bar(years,AP);
    grid on;
    xlabel('Year', 'FontSize', 12);
      %%% Decide on Y axis label %%%
        if v==1
            ylabel('Average Annual Researcher Package Estimated (*$1000)', 'FontSize',
12);
        elseif v==2
```

```
            ylabel('Average Annual Technician Package Estimated (*$1000)', 'FontSize',
12);
        elseif v==3
            ylabel('Average Annual Supporting Staff Package Estimated (*$1000)',
'FontSize', 12);
        end
    title('RnD Personnel EOM', 'FontSize', 12);
    labels={'Private Sector';'Government Sector';'Higher Education Sector';'Public RI
Sector'};
    legend(labels,'location','northwest','FontSize', 8);
    AP = reshape(AP,1,[]);
    ylim([0 max(AP)+30]);
end
toc
```