Data Visualization Assignment 3

Team Name: DV34

Data set description and remodelling applied:

The Business Dynamics Statistics (BDS) includes measures of establishment openings and closings, firm start-ups, job creation and destruction by firm size, age, and industrial sector, and several other statistics on business dynamics. The U.S. economy is comprised of over 6 million establishments with paid employees. The population of these businesses is constantly churning -- some businesses grow, others decline and yet others close. New businesses are constantly replenishing this pool. The BDS series provide annual statistics on gross job gains and losses for the entire economy and by industrial sector, state, and MSA. These data track changes in employment at the establishment level, and thus provide a picture of the dynamics underlying aggregate net employment growth.
This data set is well structured and all the columns are interconnected.
We only applied slicing of columns and rows to get different visualization and presenting the final output.
The second remodelling we did was getting mean value of different attribute to get a clear view of job creation or destruction over the period of 1978 to 2018.

1. Was the remodelling and subsequent visualization done for the entire or part of the dataset? Explain in detail the decision made here, and if a partial dataset is used, how it was determined.

   Remodelling is done on entire dataset. We have pre-processed the data and following results has been found:

Dataset statistics

Number of variables	25
Number of observations	1927
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	376.5 KiB
Average record size in memory	200.1 B

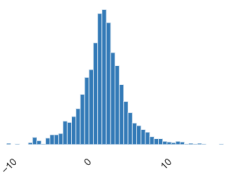
Variable types

Categorical	1
Numeric	24

Data.Calculated.Net Job Creation Rate
Real number (ℝ)

HIGH CORRELATION

Distinct	1765	Minimum	-10.405
Distinct (%)	91.6%	Maximum	17.248
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	403
Infinite (%)	0.0%	Negative (%)	20.9%
Mean	1.931559419	Memory size	15.2 KiB



Toggle details

Pre-processing results and Remodelling applied:

The given data set is Business dynamics data which has 1927 number of observation and 25 columns. This is business dynamics of different states of USA. It has 47 different states data and 41 years staring from 1978 to 2018. It also provides huge and detailed description of firm creation and destruction in different states of US. The data set is really in well format, all the required fields are there ,all thought some columns like Data. Job Creation. Count and Data. Job Destruction. Continuers etc can be generated from other fields so it is redundant.

Our main aim was to find a relation between different states of USA on the basis of how much job has been created or in other words how much employment was generated in a particular state on a specific year or overall time spam.

We have applied column wise and row wise slicing on data to get results on different visualization. For getting a clear picture of rate of employment we also found mean for job creation rate from range of years 1978 to 2018. This remodelling was for complete data.

And later we plot it on sunburst chart.

Interaction: We had data set which has job creation and destruction rate year wise. We applied mean function column wise for all states on job creation and destruction rate and later found their difference that was net job creation rate for a state.

Average of job creation rate for a state=sum of all the years data/No of years

Average of job destruction rate for a state=sum of all the years data/No of years

Net Job creation rate= Average of job creation rate for a state -Average of job destruction rate for a state

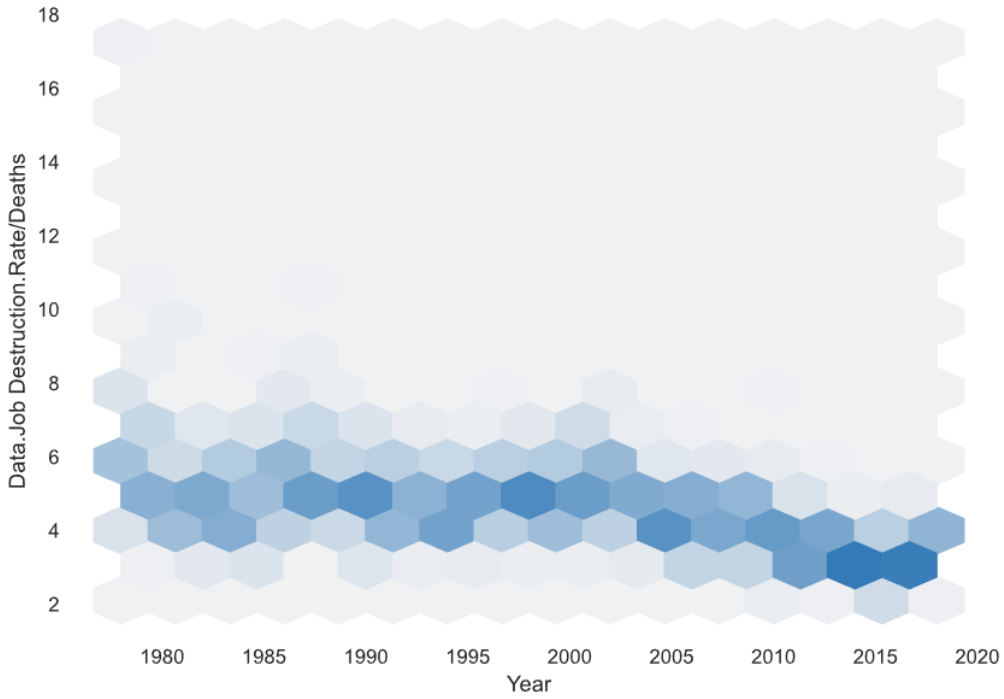


Fig 1

2. What are the inferences from each of the visualizations? What is the joint inference that can be made from all six visualizations?

Task Done

Force Directed Graph:

It is used to visualize the connections between objects in a network. By grouping the objects connected to each other in a natural way, a Force-Directed Graph is visually interesting and also makes it possible to discover subtle relationships between groups.

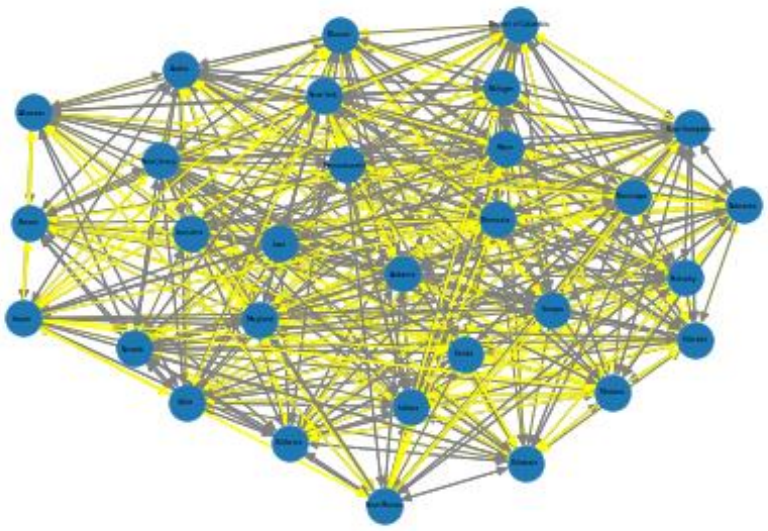


Fig 2

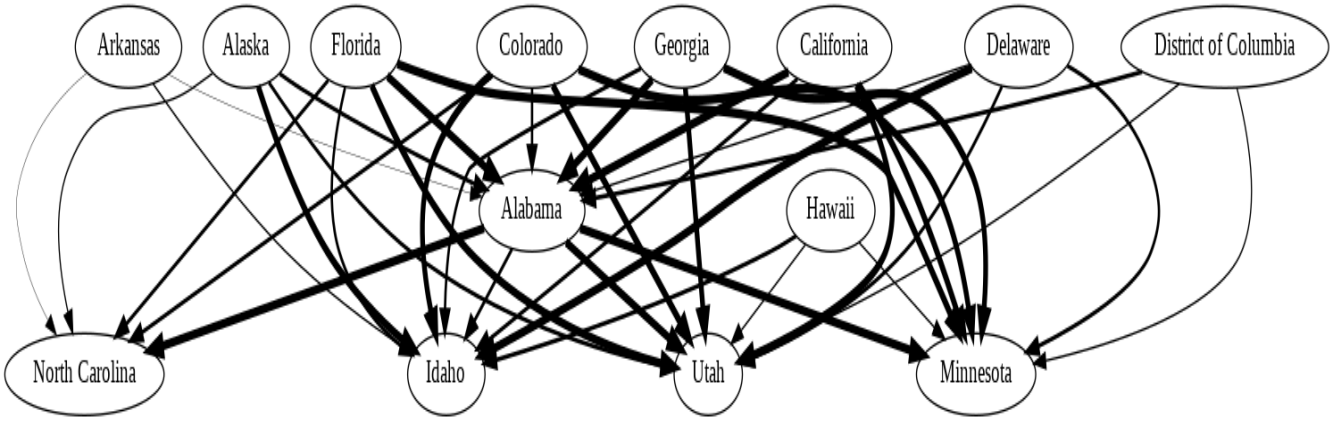


Fig 3

Tree Map:

A tree map in Python is a visualization of data that splits a rectangle into sub-parts. The size of each subpart is in proportion to the data it represents. It is somewhat like a pie-chart. Although, tree maps can represent much-more complex data as compared to a pie-chart.

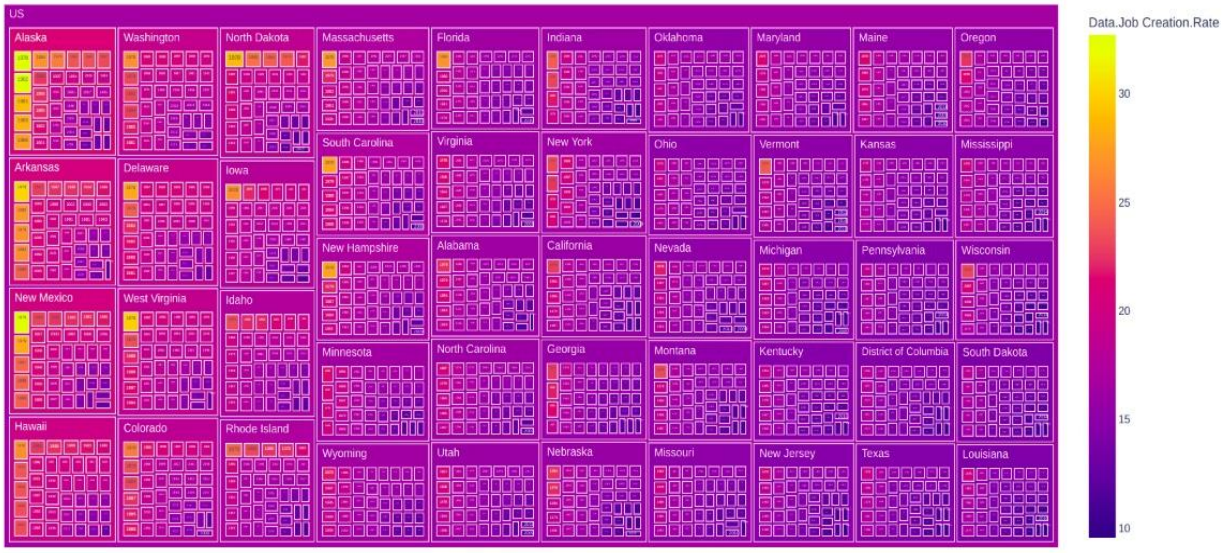


Fig 4

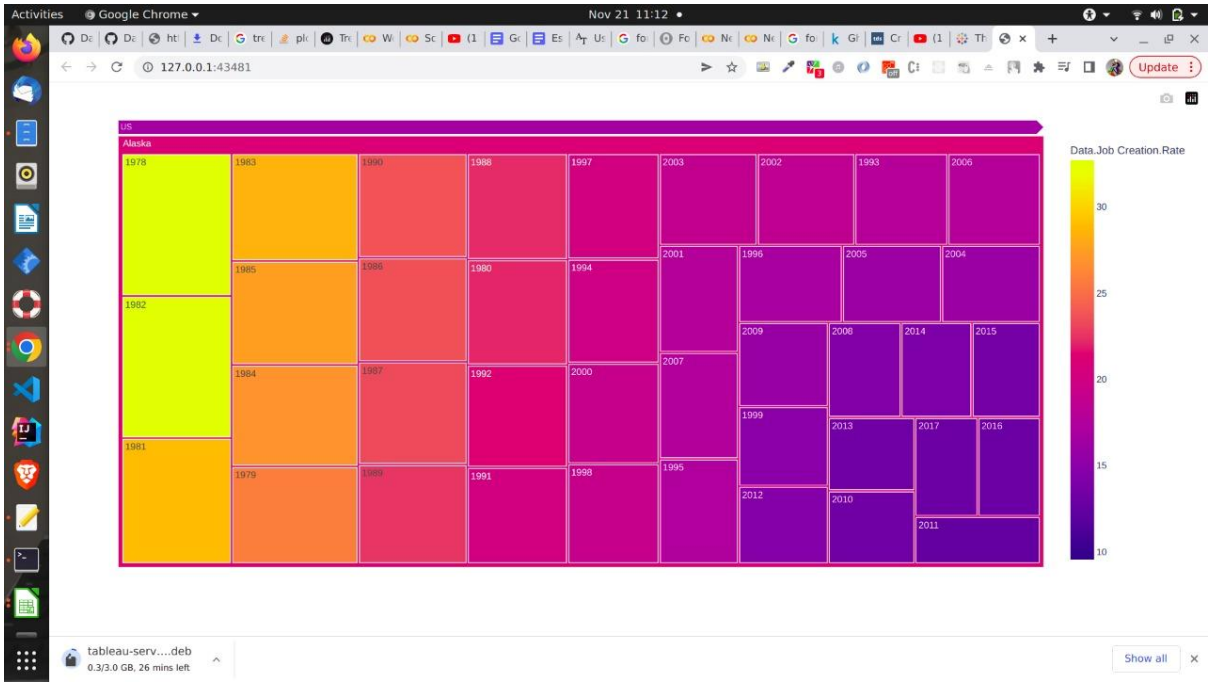


Fig 5

Inference:

This tree map shows that **Alaska has highest** job creation rate while Louisiana has lowest rate. Within Alaska 1978 has highest rate while 2011 has lowest rate.

This tree map can be used to read job creation rate of any state and particularly any year.

The colour bar represents yellow for high values while violet for low values.

Sunburst visualization

A Sunburst Diagram is used to visualize hierarchical data, depicted by concentric circles. The circle in the centre represents the root node, with the hierarchy moving outward from the centre. A segment of the inner circle bears a hierarchical relationship to those segments of the outer circle which lie within the angular sweep of the parent segment.

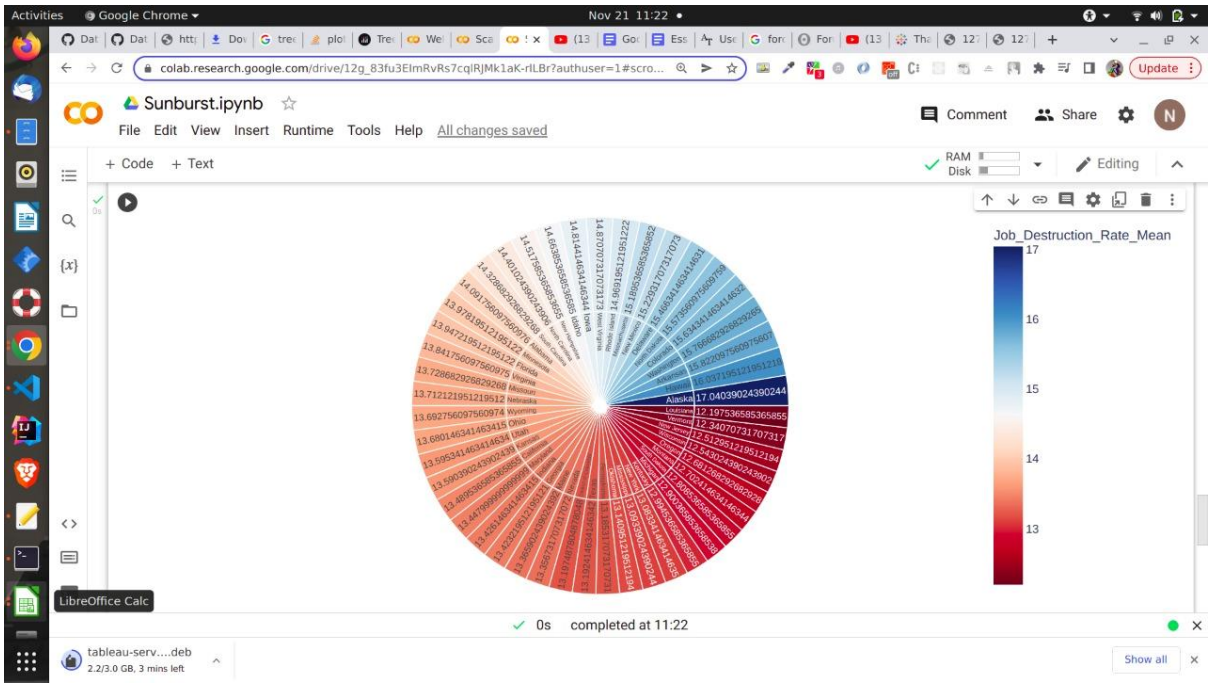


Fig 6

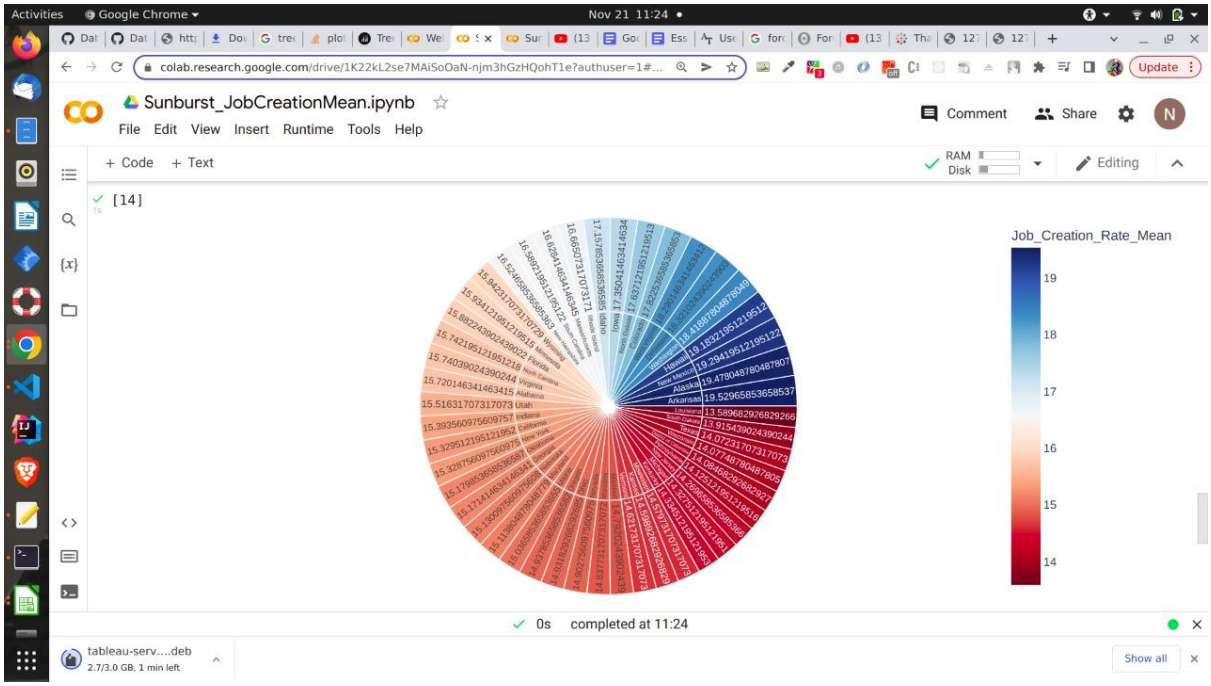


Fig 7

Inference: Sun Burst (fig 6) shows that arkansha has highest job rate creation of 19.52. Second higest is Alaska 19.47 then new jercy and Hawaii.

Sun Burst chart (fig 7) shows that Alaska has highest job destruction rate.

Job Destruction Rate

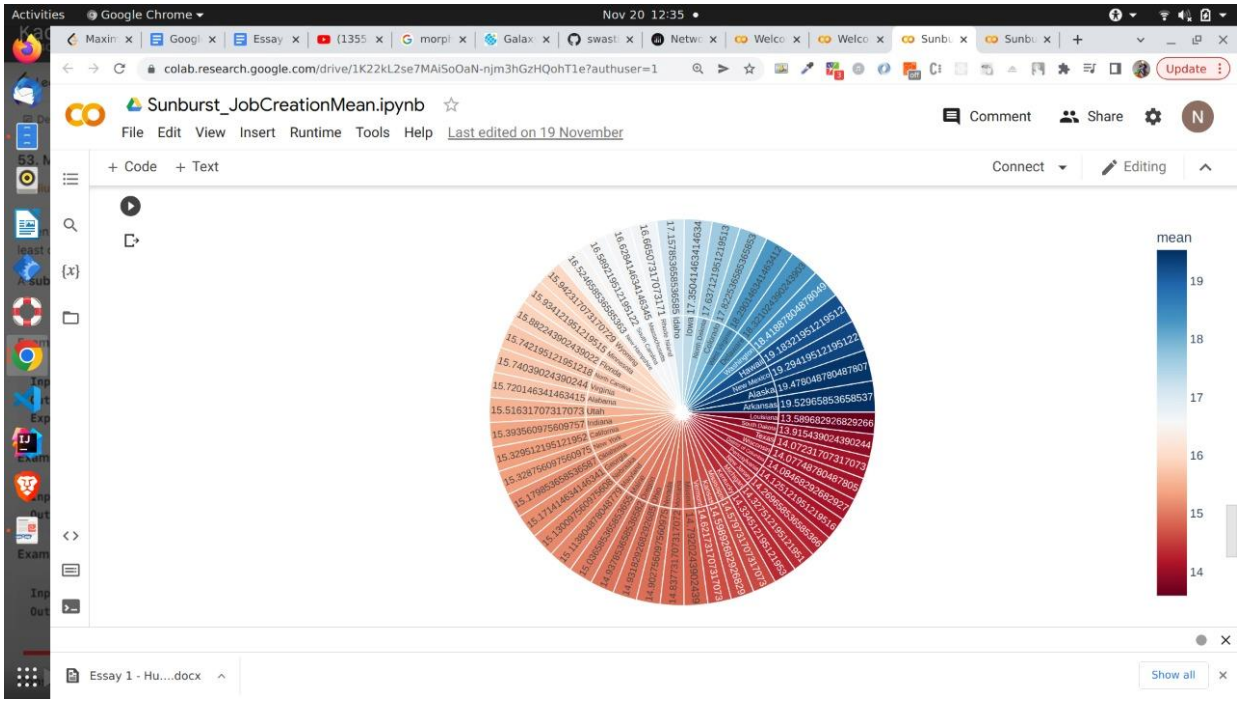


Fig 8

Parallel coordinates plot

This type of visualisation is used for plotting multivariate, numerical data. Parallel Coordinates Plots are ideal for comparing many variables together and seeing the relationships between them.

In a Parallel Coordinates Plot, each variable is given an axis and all the axes are placed parallel to each other. Each axis can have a different scale, as each variable works off a different unit of measurement, or all the axes can be normalised to keep all the scales uniform. Values are plotted as a series of lines that are connected across all the axes. This means that each line is a collection of points placed on each axis, that have all been connected.

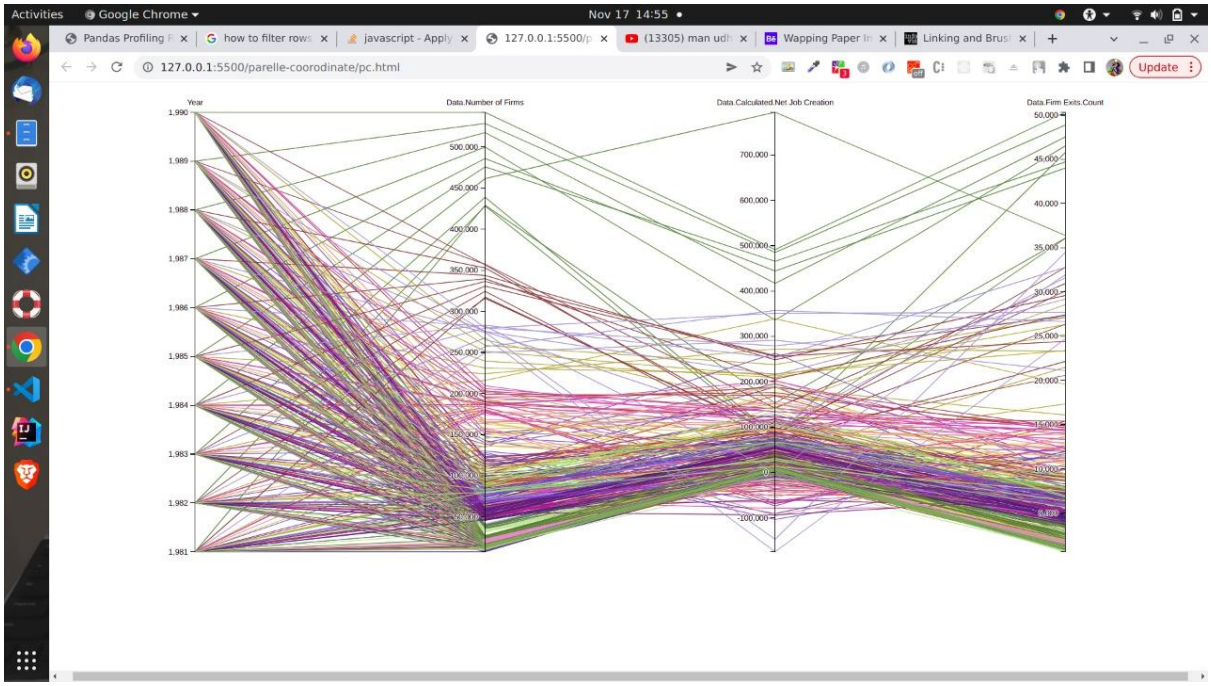

Activities
Google Chrome
Nov 17 14:55
127.0.0.1:5500/parelle-coorodinate/pc.html
Update
Year
Data.Number of Firms
Data.Calculated.Net Job Creation
Data.Firm Exits.Count

Fig 9

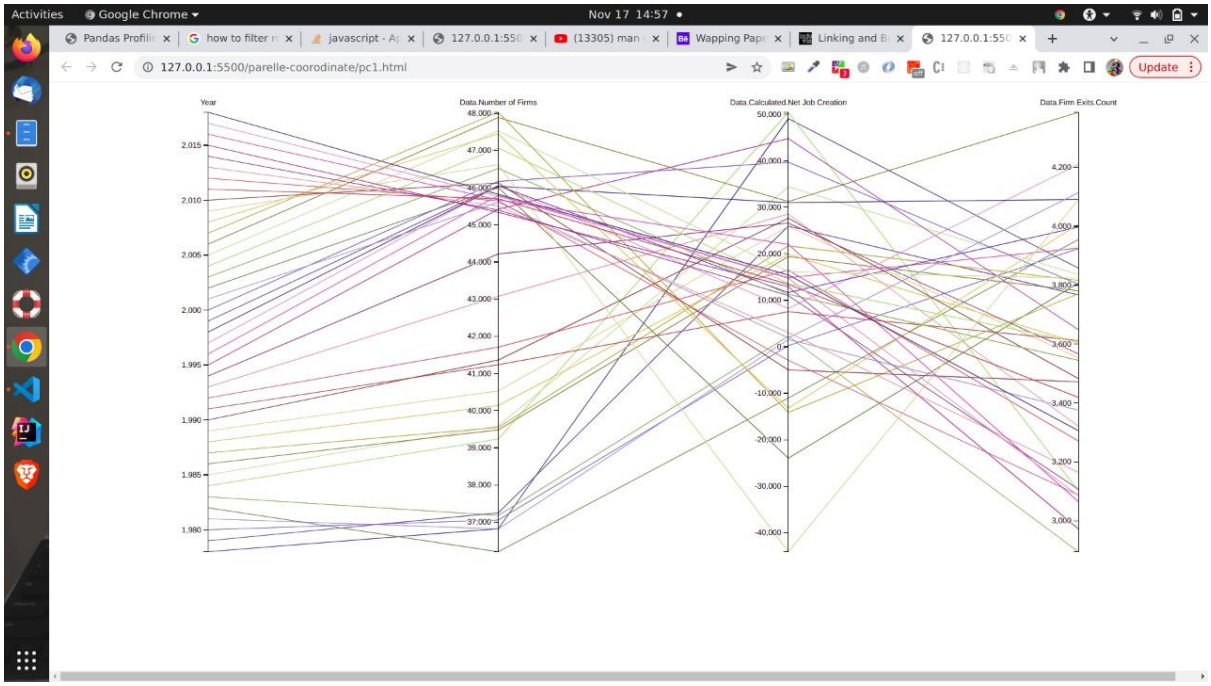

Activities
Google Chrome
Nov 17 14:57
127.0.0.1:5500/parelle-coorodinate/pc1.html
Update
Year
Data.Number of Firms
Data.Calculated.Net Job Creation
Data.Firm Exits.Count

Fig 10

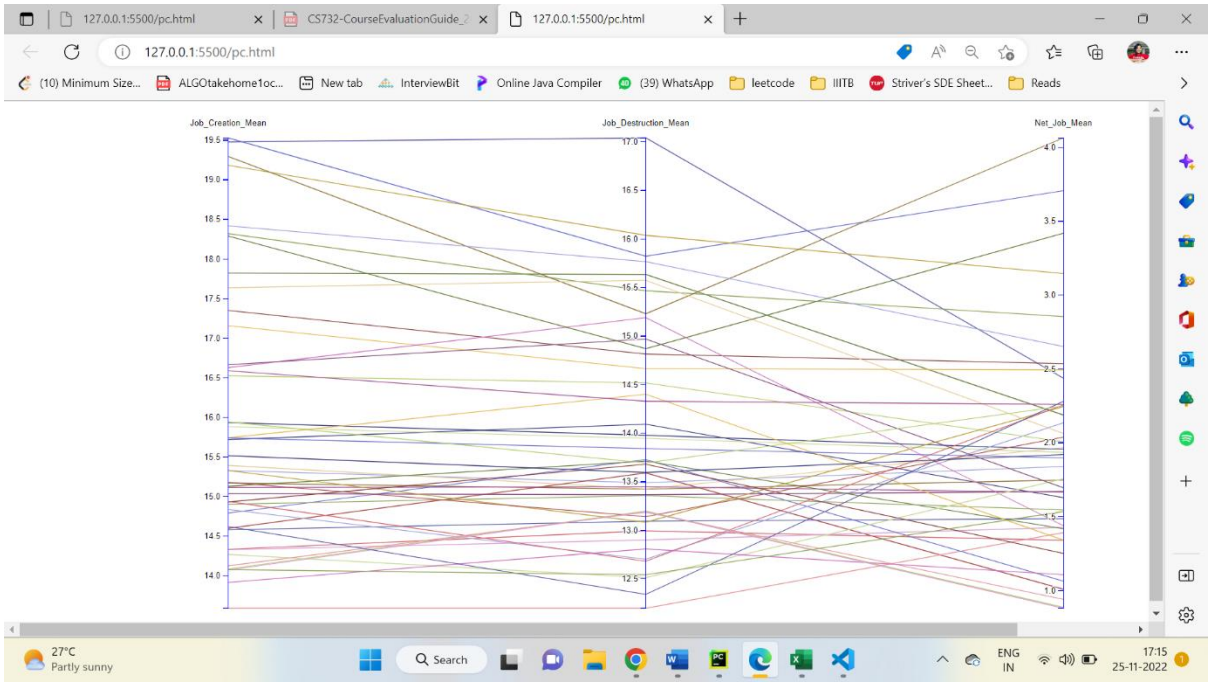


Fig 11

Inference:

Parallel coordinate plot(fig11) shows that's Alaska has highest job creation and highest job destruction rate .

Matrix visualization of the adjacency matrix:

An adjacency matrix is a way of representing a graph as a matrix of Booleans (0's and 1's). A finite graph can be represented in the form of a square matrix on a computer, where the Boolean value of the matrix indicates if there is a direct path between two vertices.

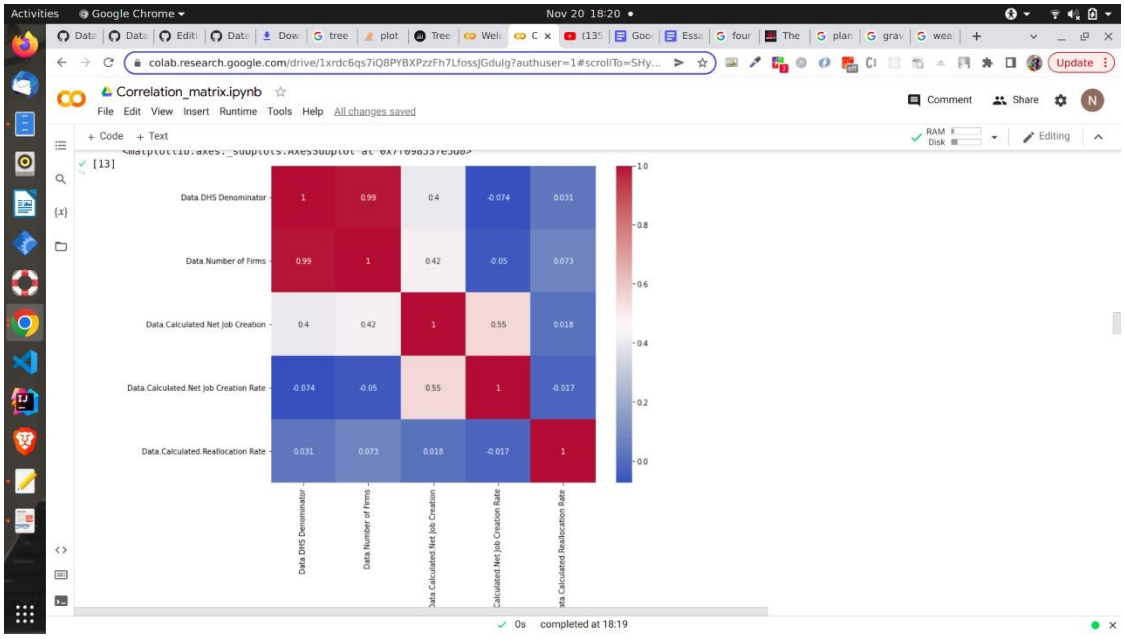

Activities
Google Chrome
Nov 20 18:20
colab.research.google.com/drive/1xrdc6qs7iQ8PYBXPzzFh7LfossjGdulg?authuser=1#scrollTo=SHy...
Update
Correlation_matrix.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
[13]
Data.DHS Denominator
Data.Number of Firms
Data.Calculated.Net Job Creation
Data.Calculated.Net Job Creation Rate
Data.Calculated.Reallocation Rate
1 0.99 0.4 -0.074 0.031
0.99 1 0.42 -0.05 0.073
0.4 0.42 1 0.55 0.018
-0.074 -0.05 0.55 1 -0.017
0.031 0.073 0.018 -0.017 1
1.0
0.8
0.6
0.4
0.2
0.0
0s completed at 18:19

Fig 13

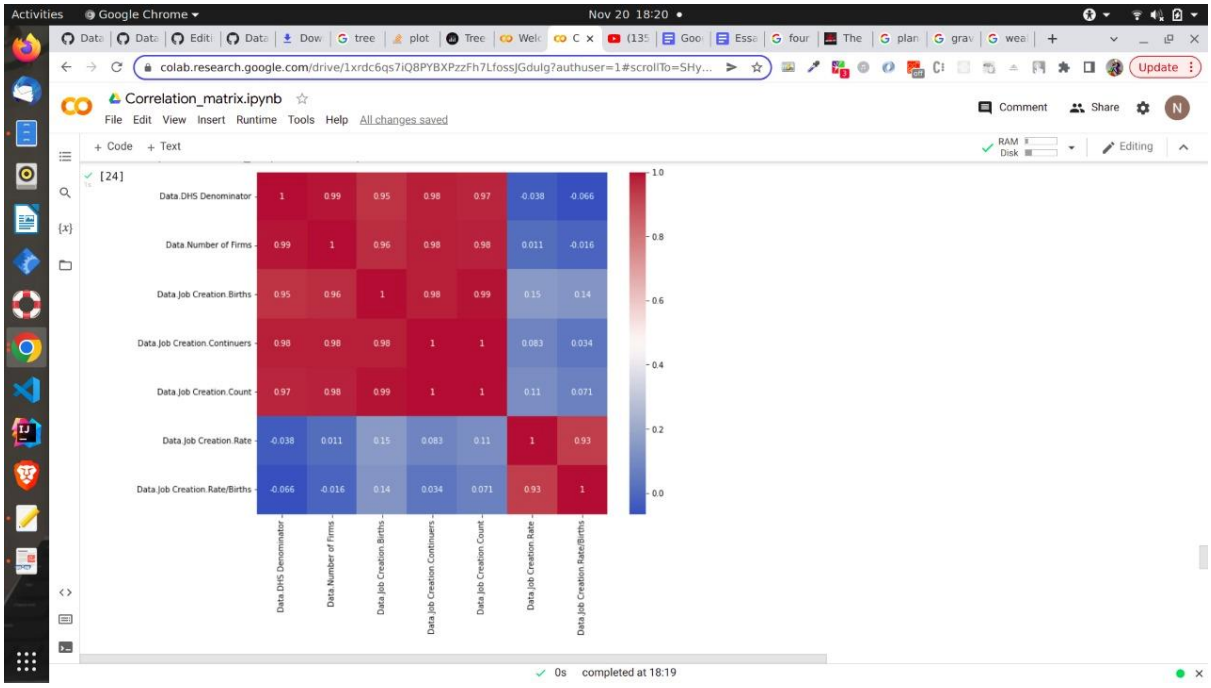

Activities
Google Chrome
Nov 20 18:20
colab.research.google.com/drive/1xrdc6qs7iQ8PYBXPzzFh7LfossjGdulg?authuser=1#scrollTo=SHy...
Update
Correlation_matrix.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
[24]
Data.DHS Denominator
Data.Number of Firms
Data.Job Creation.Births
Data.Job Creation.Continuers
Data.Job Creation.Count
Data.Job Creation.Rate
Data.Job Creation.Rate/Births
1 0.99 0.95 0.98 0.97 -0.038 -0.066
0.99 1 0.96 0.98 0.98 0.011 -0.016
0.95 0.96 1 0.98 0.99 0.15 0.14
0.98 0.98 0.98 1 1 0.083 0.034
0.97 0.98 0.99 1 1 0.11 0.071
-0.038 0.011 0.15 0.083 0.11 1 0.93
-0.066 -0.016 0.14 0.034 0.071 0.93 1
1.0
0.8
0.6
0.4
0.2
0.0
0s completed at 18:19

Fig 14

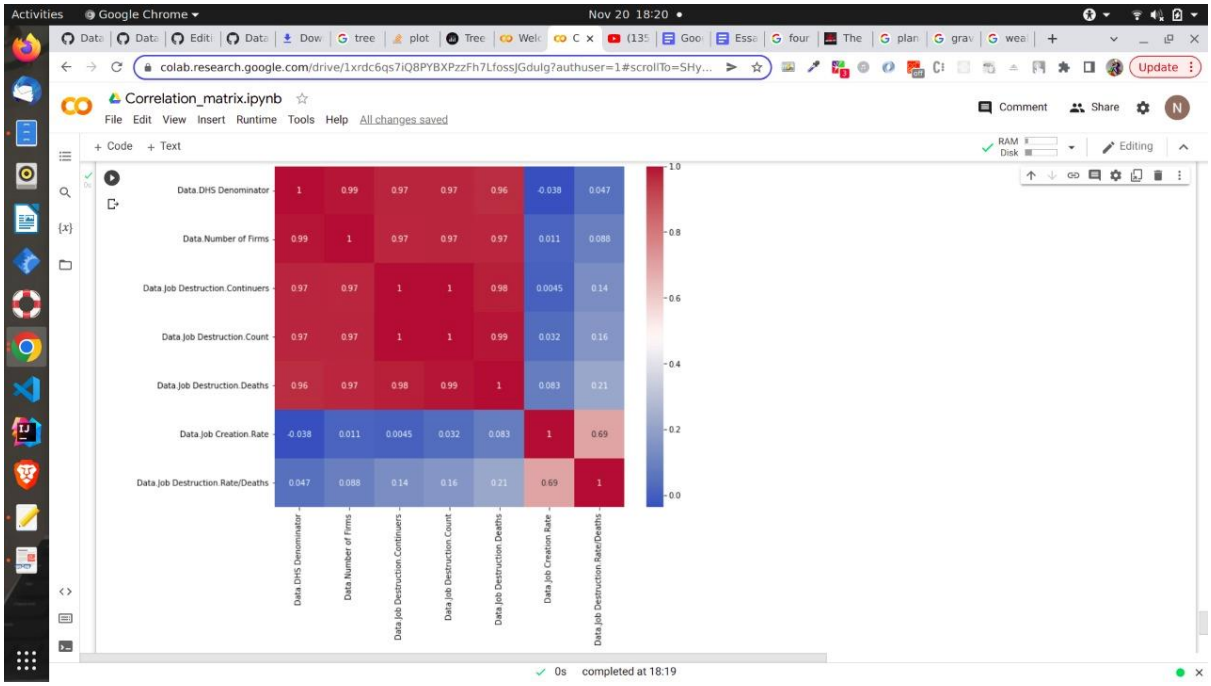


Fig 15

Inferences:

The corelation matrix show that ther is corelation between columns Number_of_firms and DHS_denominator.

Also in job_creation_birth, job_creation_continuer and job_creation_count there is high corelation.

Job_destruction_deaths, job_destruction_continuer and job_distruction_count have high corelation

Scatterplot matrices:

Scatter plots are used to observe relationship between variables and uses dots to represent the relationship between them. The **scatter()** method in the matplotlib library is used to draw a scatter plot. Scatter plots are widely used to represent relation among variables and how change in one affects the other.

matplotlib.pyplot.scatter(x_axis_data, y_axis_data, s=None, c=None, marker=None, cmap=None, vmin=None, vmax=None, alpha=None, linewidths=None, edgecolors=None)

The scatter() method takes in the following parameters:

- **x_axis_data-** An array containing x-axis data
- **y_axis_data-** An array containing y-axis data
- **s-** marker size (can be scalar or array of size equal to size of x or y)
- **c-** color of sequence of colors for markers
- marker- marker style

Fig 16

Scatter Plot:

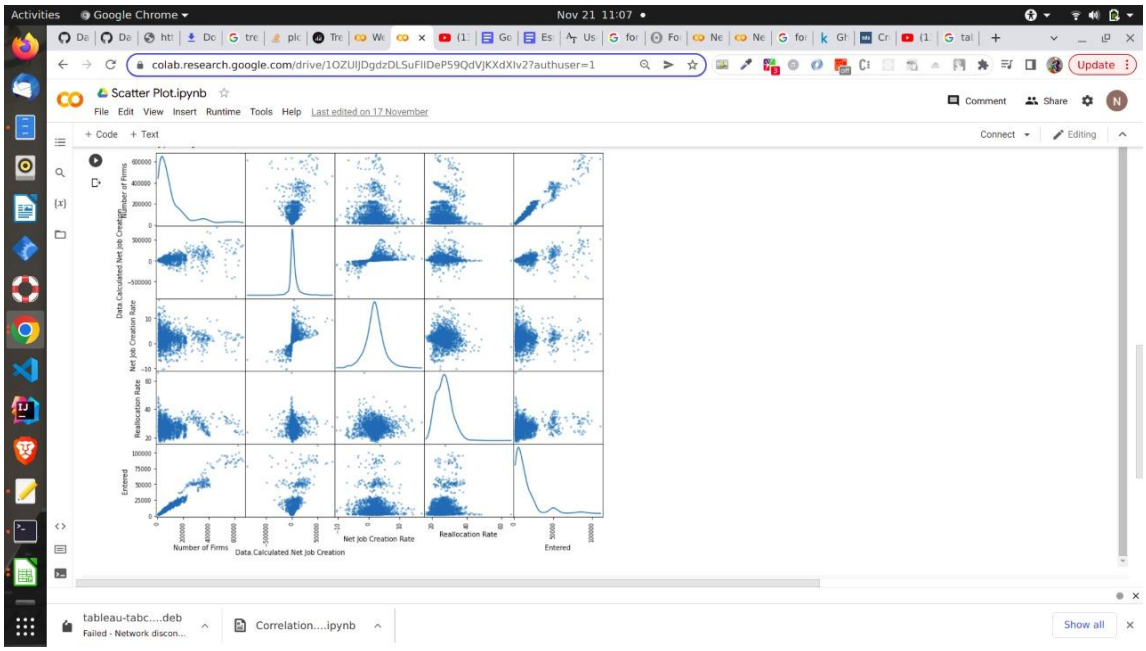


Fig 17

Inference: here we can see denser ellipses forming at the intersection of columns Net Job Creation and Reallocation Rate, which reveals that as the reallocation increases more jobs are getting created.

Conclusion: From all the

1)Inferences got from different visualization we got end result as Alaska has highest job creation as well as highest job destruction rate.

2) Arkansas has highest job creation.

3) Year 1978 was best year to get job while 2009 had worst year for people.

4) As the reallocation increases job creation increases.

Work Done:

Pragya: MT2021094: Force directed graph, Tree Map, Sun burst charts

Nikunj: MT2021084: Scatter plot matrix, parallel coordinate, Matrix Visulization