CS7DS4 / CSU44065 Data Visualization 2019-20 Assignment A3

Student Name : Ashish Arvindrao Kannur

Student No: 19300875

Declaration: "I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism ‘Ready Steady Write’, located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write."

A study on STD Infections

Background: STDs are the diseases that are transmitted sexually from one person to another. Some of their types are Chlamydia, herpes, HPV, Crabs, HBV, Gonorrhea, Syphillis, etc. Untreated STDs lead to cancerous and other terminal diseases. Epidemic of STD lasted from 1996 until 2008, after which researchers found cures to it. These visualisations give an entire overview of the disease scenario over these years.

Tasks:

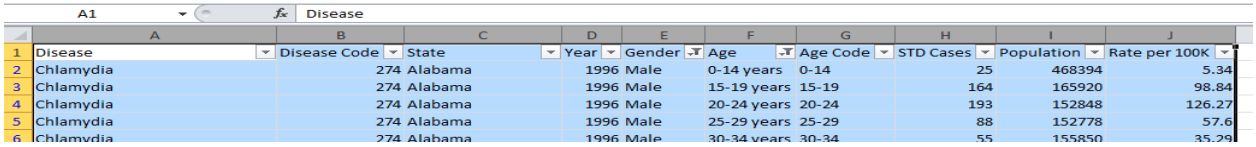
- These visualisations simplistically and effectively represent the spread of different types of STD diseases over the years.
- The patterns of STDs are analysed with respect to number of cases each year, infection rates every year, % change in the infection rates over the years, spread in terms of rate of infection and number of cases across different age groups every year, spread across different states and which regions had higher count of cases or had higher rates of infections, female and male cases, etc.

Visualisation Tool: Tableau

Data description: Here, 2 datasets which are related and which hold records of STD cases i.e. Dataset1 and Dataset2 are taken. Dataset 1 is chosen for exploratory visualisation and Dataset 2 is chosen for explanatory visualisation.

Dataset 1: This dataset gives an idea about the Distribution of the 3 types of the STDs(Chlamydia, Gonorrhea, Syphillis) across different states, ages, genders, years alongwith the number of cases, population of the respective state and Rate per 100K people.

- There are total 42631 records related to 3 types of STDs in the original dataset taken.
- Number of features or fields in the dataset are 10 namely: Disease, Disease code, State, Year, Gender, Age, Age code, STD cases, Population, Rate per 100k. Below is a small snapshot of the Dataset 1.



	A	B	C	D	E	F	G	H	I	J
1	Disease	Disease Code	State	Year	Gender	Age	Age Code	STD Cases	Population	Rate per 100K
2	Chlamydia	274	Alabama	1996	Male	0-14 years	0-14	25	468394	5.34
3	Chlamydia	274	Alabama	1996	Male	15-19 years	15-19	164	165920	98.84
4	Chlamydia	274	Alabama	1996	Male	20-24 years	20-24	193	152848	126.27
5	Chlamydia	274	Alabama	1996	Male	25-29 years	25-29	88	152778	57.6
6	Chlamydia	274	Alabama	1996	Male	30-34 years	30-34	55	155850	35.29

[1]Link for dataset: https://data.world/makeovermonday/2019w31

Pre-Processing of Dataset 1: First, null values (around 20000 rows missing values) were dropped and then redundant columns like Disease code and Age were also deleted. Data after 2008 were dropped as scientists found cure after 2008. Some new calculated fields were created for effective visualisations. The newly calculated fields are as below (These fields were calculated and included directly in the Tableau worksheets):

Calculated field	Formula
Infection Rate	sum ([STD Cases])/sum([Population.]) * 100000
Int Population	INT(Population)
Rate of Infection Rate	(ZN([Infection Rate]) - LOOKUP(ZN([Infection Rate]), -1)) / ABS(LOOKUP(ZN([Infection Rate]), -1))

Final fields considered for visualising: Disease, State, Year, Gender, Age Code, STD cases, Population, Rate per 100K and the 3 calculated fields. The total records left after dropping the nulls and records after 2008 were 21435.

Dataset 2: This dataset depicts information about 6 categories of STDs across all states, genders and ages.

- There are total 50001 records showing state wise, gender wise and age wise information about 8 different STDs.
- The columns and fields of this dataset are: state, std, gender, age.

[2]Link to the dataset: https://www.kaggle.com/bobnis/us-stats-std

	state	std	gender	age
2	Wyoming	Gonorrhea	f	30
3	Utah	Gonorrhea	m	56
4	Wyoming	Gonorrhea	f	68

Pre-processing on Dataset 2: Pivot tables were calculated as shown below (Pivot tables and its corresponding data sheets):
Male and female datasets were separated from below Pivot table 1. Additional datasheet of 'spread of STD on world map' was created by taking 2 fields from Pivot Table 1(State and Grand Total).

Pivot table 1:

Count of std	Column Labels															
	Chlamydia		Chlamydia Total	Crabs		Crabs Total	Gonorrhea		Gonorrhea Total	HBV		HBV Total	Herpes		Herpes To
Row Labels	f	m		f	m		f	m		f	m		f	m	
Alabama	1324	1392	2716	161	169	330	1306	1373	2679	93	115	208	96	100	
Alaska	1376	1267	2643	173	132	305	1367	1376	2743	107	110	217	110	105	
Arizona	1297	1318	2615	148	177	325	1325	1349	2674	99	106	205	105	123	
Arkansas	1324	1330	2654	181	164	345	1330	1353	2683	106	105	211	109	110	

Pivot Table 2:

Count of std	Column Labels								
Row Labels	Chlamydia	Crabs	Gonorrhea	HBV	Herpes	HPV	Other	Syphillis	Grand Total
Alabama	2716	330	2679	208	196	2665	447	862	10103
16	36	6	50	5	3	41	4	10	155
17	50	4	35	4	4	36	4	8	145

'Disease Spread as per age' was created using this pivot table.

The Insightful Visualisations: This section describes the dashboards created.

Approach for dashboards: First of all, as described above in the pre-processing of data section, the data is explored by taking into consideration useful columns and records, new fields are created by combining two or more columns and arrangement of data is also changed using pivot tables to create meaningful datasets. Two parameters i.e. "Infection Rate" and "Rate of Infection Rate" are created for showing Infection rates and changes in them across all the years.

In the visualisations, Consistent colour encoding is used for different types of STDs. For interactive visualisations, common filters of Disease, State, Year, Age for all charts is provided to view all visualisations by selecting filter value at single location. E.g. when Alaska is selected, then Alaska data is highlighted in several charts that allow a state selection. Forecast indicator is included in Infection rate plot to predict future value based on previous years' trend. Light coloured bars indicate the future estimated values.

Line plots are used where trend is intended to be shown along with the values. Bar plots are used to show numerical quantities like number of cases, infection rates, population, etc with respect to categorical values like year, age and state. Stacked bar plots are used to show multiple categories of diseases along with their numerical values. Finally Maps are used to show the regions those were affected. Darker shades in all the plots represent the higher values whereas lighter ones denote lower values. The scale of colour shades on the plots provide the range of values these shades represent. Hovering over the plots or maps will give the exact values of the parameters involved in the graph at different points.

Dashboard 1 (Dataset 1) (Exploratory Analysis):

- **State wise Rate of Infection**: Geographical representation of Rate of Infection in different states on a map. The radial sizes and shades of the circles vary with the values of "Rate of Infection".
- **Distribution of Population and Rate of 3 diseases per 100K people**: Stacked Bar plot shows the state wise and year wise populations belonging to each category of STD while the Line plot shows "Rate per 100K" info of the infection for each state . The Line plots with different colours show the Rate per 100K for 3 different STDs for each state for all the years. These plots can be used to compare the values of populations of and Rates among different states. The line plot also depicts the trend in the rate for each individual disease over the years.
- **State wise Spread of Diseases**: Horizontal Bar plot represents rates of infection for different states and can be used to compare state wise Rate of Infection for each of the 3 STD diseases. There are 2 categorical(State and Disease) and 1 quantitative(Rate of Infection)

Dashboard 2 (Dataset 1) (Exploratory Analysis):

- **Rate of Infection**: A Bar plot here represents yearwise representation of the calculated field "Infection Rate". Here the categorical attribute is Year and quantitative one is "Rate of Infection". An additional feature of forecasting future values is also implemented.
- **Yearwise % change in the Infection Rate**: Bar plot showing the % change in the Infection rate every between every 2 consecutive years. Darker shade represents higher value. Here one categorical (Year) and 1 quantitative(Rate of Infection Rate) attributes are used. This plot gives an idea about the pattern of Infection Rate over the years.
- **Rate of Infection and its % change over years**: A combined plot of "Rate of Infection" (Bar plot) and "Yearwise % change in the Infection Rate"(Line plot). This combined plot gives the pattern of infection rate over the years along with the exact values of increase and decrease in percentage.
- **Yearwise Rate per 100K across all age ranges**: Horizontal bar graph showing "Rate per 100K" for each age group over the years. This plot helps visualise and compare the pattern and values of Rate of infection varying across all age groups. The categorical attributes here are Age and Year while the quantitative attribute is Rate per 100K.
- **Yearwise Number of STD cases across all age ranges**: Another horizontal bar plot showing yearwise number of 'STD cases' for all age ranges. This plot helps visualise the pattern of number of STD cases varying across all age groups over the years. The categorical attributes here are Age and Year while the quantitative attribute is STD cases.

Dashboard 3 (Dataset 2)(Explanatory Analysis):

- **Total STD cases**: This is a Geographical data representation of Number of STD cases. The colour shades of states vary according to the Number of STD cases.
- **Female STD cases**: Stacked Bar plot showing statewise representation of number of females affected by different types of STDs. Different colours are used to represent different categories of STDs. Bars are separated by States and each bar is subdivided into 8 categories of STDs. Eight quantitative attributes (counts of Chlamydia/Crabs/ Gonorrhea/HBV/Herpes/HPV /Other/Syphillis) and one categorical value(State) are used here.
- **Male STD cases**: Stacked bar plot showing statewise representation of number of males affected by different types of STDs. Different colours are used to represent different categories of STDs. Bars are separated by States and each bar is subdivided into 8 categories of STDs. Here 8 quantitative attributes (counts of Chlamydia/Crabs/Gonorrhea /HBV/Herpes/ HPV/Other/Syphillis) and one categorical value(State) are used.
- **Number Of Cases of Different STDs in each age group for each state:** Stacked bar plot showing statewise Number Of Cases of Different STDs in each age group. Different colours are used to represent different categories of STDs. Bars are separated by Age and each bar is subdivided into 8 categories of STDs .Here 8 quantitative attributes (counts of Chlamydia/Crabs/ Gonorrhea/HBV/Herpes/HPV/Other/Syphillis) and 2 categorical values(State and Age) are used.

Conclusion:

- From the above mentioned visualisations it was seen that District of Columbia had low population count but had highest Overall as well as individual Disease wise rate of infection.
- Other states that recorded highest rate of infections were Alaska, Mississippi, South Carolina and Louisiana. These are clearly visible in Dashboard 1.
- From Dashboard 2, the rate of infection increased from 1996 to 2008 each year and highest rate was in 2008.
- The rate of infection changed more rapidly in the year 1998 w.r.t year 1997 as compared to all other years (Increase by 12%)
- People between ages of 15 to 30 years showed were affected the most by STDs. The rate per 100K people and number of cases belonging to this age group were highest. Especially the 15-19 and 20-24 age groups showed highest numbers from around 2002 to 2008.
- From Dashboard 3, it can be seen that California, Colorado, Illinois and Pennsylvania were among the states that recorded highest number of total STD cases.
- Male cases were in more in number as compared to female cases.

References:

[1] Kriebel@VizWizBI, A. (2019).data.world. Retrieved April 2020, from data.world: https://data.world/makeovermonday/2019w31/

[2] B. Nis, "kaggle.com," kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/bobnis/us-stats-std. [Accessed April 2020].