IS6030

Data Management
HW5 Submission

Swapnil Sharma
M10743959

Description of Data:

Dataset was downloaded from url: https://www.dhs.gov/immigration-statistics/yearbook/2014

From Lawful Permanent residents 2014 (Tables 1-12) 4 tables were used for the project. They were Table 3d, Table 4, Table 11d, Table 12d.

There are 4 tables in the data set and all of them are independent.

- Table 3d contains number of PERSONS OBTAINING LAWFUL PERMANENT RESIDENT STATUS BY REGION AND COUNTRY OF BIRTH: FISCAL YEARS 2005 TO 2014
- Table 4contains number of PERSONS OBTAINING LAWFUL PERMANENT RESIDENT STATUS BY STATE OR TERRITORY OF RESIDENCE: FISCAL YEARS 2005 TO 2014
- Table 11d contains number of PERSONS OBTAINING LAWFUL PERMANENT RESIDENT STATUS BY BROAD CLASS OF ADMISSION AND REGION AND COUNTRY OF LAST RESIDENCE: FISCAL YEAR 2014
- Table 12d contains number of IMMIGRANT ORPHANS ADOPTED BY U.S. CITIZENS BY SEX, AGE, AND REGION AND COUNTRY OF BIRTH: FISCAL YEAR 2014

Overview of columns and Data Stored:

All tables were in .xls format.

- Country: Shows the name of the country of origin of the person obtaining PR. Text values.
- State: Shows the name of state of USA. Text values.
- Year 2005-2014: It Captures No. of Person. Integer Values.

Normalization of Data:

The data in all the Tables is of the normal form. There is one row of every country or state and columns capture number of persons yearwise.

Problems in the Data:

- As the Dataset contains sensitive information there were cells where data was not available and value stored as NA, data was not applicable and value stored as X, data withheld to limit disclosure and value stored as D and '-' was used to store 0.
- I created one new excel file to clean the data and replace all the values which were not stored as text as null values while replaced '-' with 0.

- Column names were not adhering to best practices and hence I changed them.
- Column of Total was removed from Dataset as it was inconsistent due to limit disclosure data breakdown for certain fields. Total is calculated in SQL for study based on disclosed data.

Data exploration using SQL:

- New file with refined dataset was migrated to SQL to further explore it.
- Table names were defined to make it more meaningful.
- People from over 208 countries were given Permanent Resident (PR) between year 2005-2014.
- Following table represent statistics of total number of person getting PR over the Years.

Total PR in 2005	Total PR in 2006	Total PR in 2007	Total PR in 2008	Total PR in 2009	Total PR in 2010	Total PR in 2011	Total PR in 2012	Total PR in 2013	Total PR in 2014
1122257	1266123	1052415	1107121	1130813	1042625	1062035	1031627	990549	1016518

- Using a complex SQL query it was found out in which year max number of PR was given to residents of which country. Following was the founding.

	country	totalpr	year
1	Mexico	189989	2008

- Analysis was further drill down to year 2014 and top 6 countries by number of persons who got PR was found.

	country	Year_2014
1	Mexico	134052
2	India	77908
3	China, People's Republic	76089
4	Philippines	49996
5	Cuba	46679
6	Dominican Republic	44577

- From State table I made attempt to find on an average which were the top 10 states to host PR. Following table represents the findings that state with centers of Business like LA, New York city, Austin, Atlanta etc host maximum PR.

	state	Average_PR_Granted
1	California	219779
2	New York	146905
3	Florida	119743
4	Texas	91321
5	New Jersey	55877
6	Illinois	41858
7	Massachusetts	31728
8	Virginia	29628
9	Georgia	27341
10	Maryland	25845

- I tried to summarize the classification of PR given for top 6 countries in 2014.Following table summarizes the data.

	COUNTRY	Familysponsored	Employmentbased	ImmediaterelativesofUScitizens	Diversity	Refugees	Other
1	China, People's Republic	15867	20783	23959	30	11681	172
2	Cuba	3676	4	2524	279	40008	14
3	Dominican Republic	25001	298	18988	6	117	140
4	India	15740	38593	18480	84	909	645
5	Mexico	33976	7292	81347	19	646	9827
6	Philippines	16476	7740	24207	NULL	NULL	174

- Following table summarizes the number of persons who got adopted in 2014 were of which country of origin.(TOP 10)

	country	TotalAdoption
1	China, People's Republic	2002
2	Ethiopia	681
3	Haiti	414
4	Ukraine	409
5	Korea, South	373
6	All other countries	218
7	Uganda	200
8	Bulgaria	180
9	Philippines	167
10	Colombia	155

- Following table represent females of which country origin were adopted maximum.

	country	PercentFemaleAdopted
1	India	71.64
2	Hungary	68.18
3	Guyana	66.67
4	Pakistan	64.71
5	Saint Vincent and the Grenadines	63.64
6	Latvia	60.66
7	Jamaica	59.18
8	China, People's Republic	58.89
9	Marshall Islands	58.62
10	Colombia	58.06

Statistics Reporting using SAS:

As SAS is one of the best sophisticated tool available for statistical analysis, I gathered all relevant data using PROC Univariate command on it. Here are some findings of same:

- For year 2014

Basic Statistical Measures			
Location		Variability	
Mean	4887.106	Std Deviation	13523
Median	825.000	Variance	182868915
Mode	5.000	Range	134052
		Interquartile Range	3647

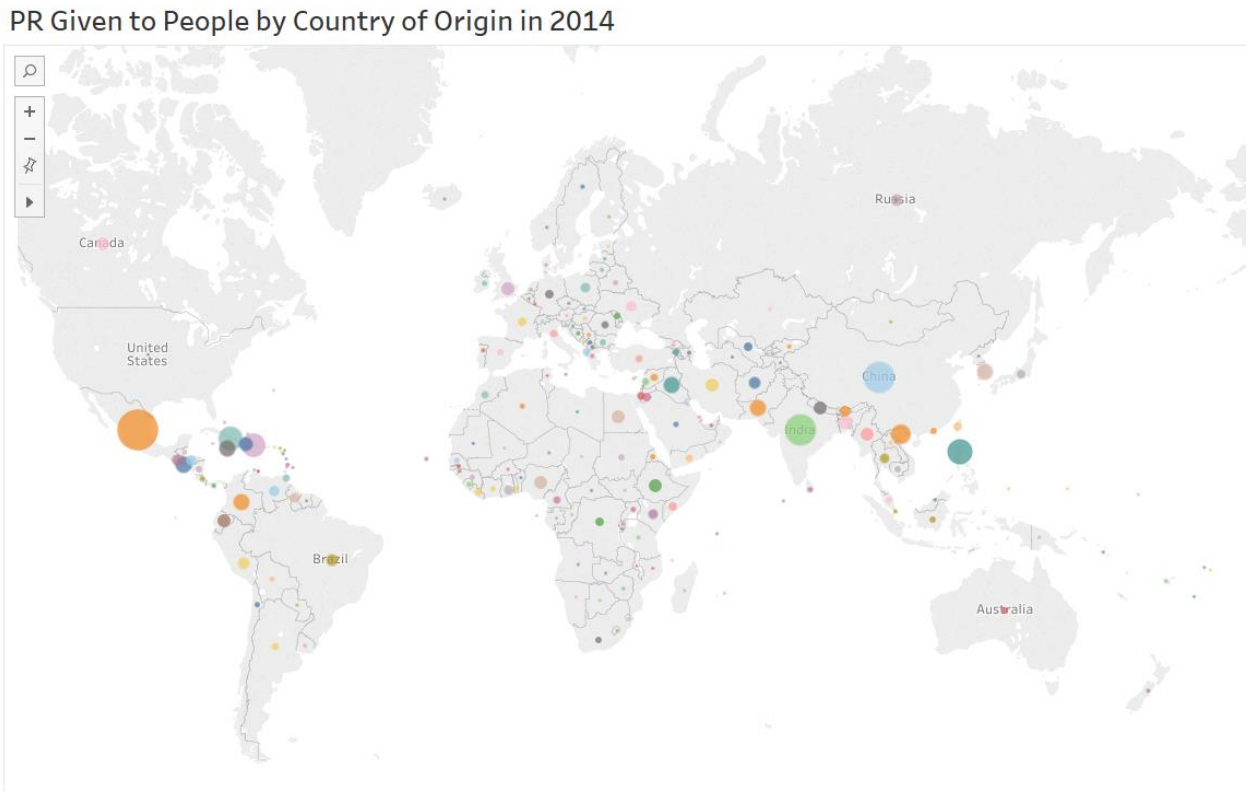
- It clearly shows that data is skewed with such large range and thus people from certain countries are only getting PR and it is not equally distributed.
- For Mexico, person from this country got max PR across the decade of 2005-2014, are the stats summarized below.

The UNIVARIATE Procedure
Variable: Mexico (Mexico)

Basic Statistical Measures			
Location		Variability	
Mean	153679.5	Std Deviation	18366
Median	147523.0	Variance	337326398
Mode	.	Range	55937
		Interquartile Range	25800

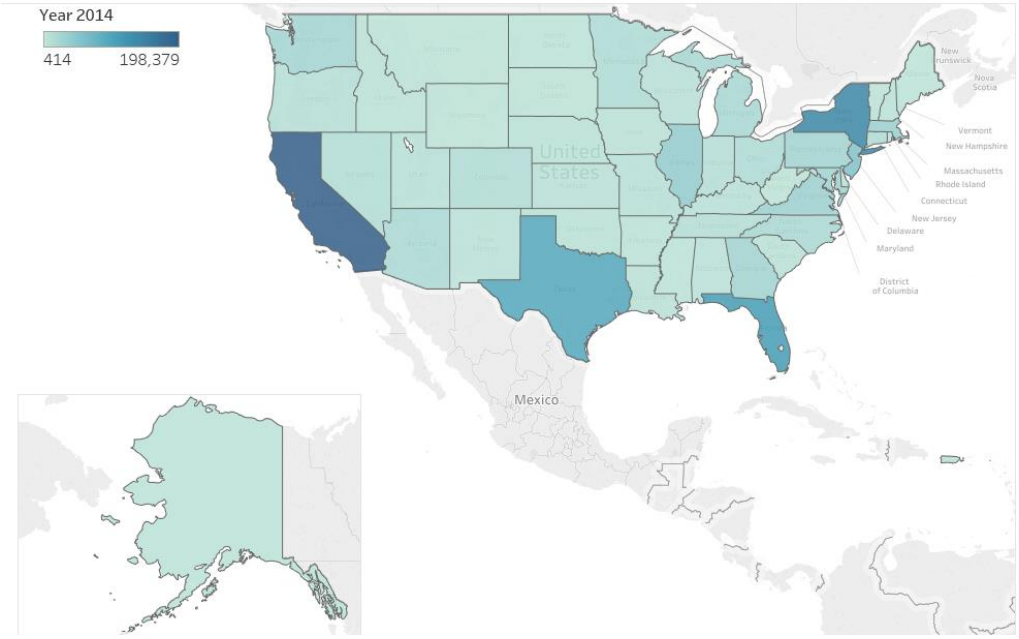
- This shows that Mexico origin people have been consistently getting PR since 2005 till 2014.

Tableau for Data Visualization:



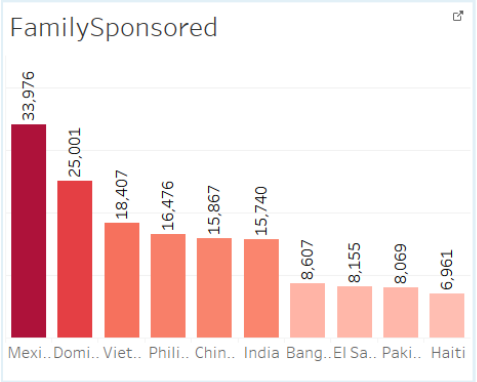
- Snapshot of world where size of bubble indicates the number of person getting PR in USA from that particular Country.

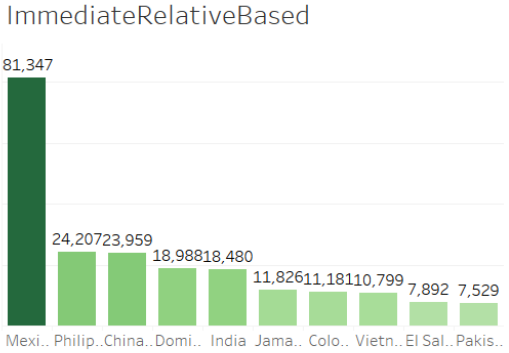
State Giving PR in 2014

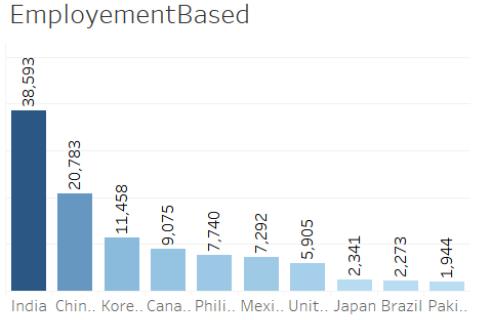


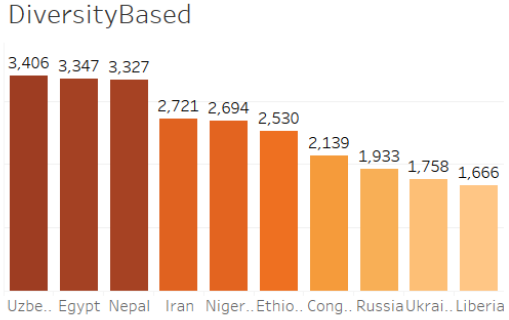
- Above visualization shows which states hosted PR in 2014. Sequential use of color is done to indicate number.

PR given to people based on Category varies across OriginCountry









Summary of Findings:

- People from all over the world (across 200 + countries) migrate to USA and they have been given PR over the years. Though the migration from certain countries is in large numbers compared to others.
- Neighbouring Countries like that of Mexico, Canada and Asian countries like that of India, china,South Korea and Philipines are the major one to get max PR in 2014.
- Mexico origin people got maximum PR across 2005-14 with max in 2008.
- States with big centers of Business like that of LA, New York city, Atlanta, Chicago,Austin etc were the major host of people getting PR.
- Dashboard shows though Mexico origin people got Maximum PR but it was due to their Family or Immediate relatives sponsoring them but Indian and Chinese origin people got due to their requirement in workforce via employers sponsoring them.

Challeneges Faced:

- Cleaning of data set by changing null values, renaming columns, Tables etc before migrating it to each of SAS, SQL and Tableau were time consuming.
- I discvered that Tableau cannot be used to make a dashboard with different tables in database with no joins in between them. Hence three different Tableau files were made for this project.
- I was not able to get statistic summary in SAS using PROC Transpose command and hence I Pivoted the table in Excel itself and then did PROC Univariate to get country wise summary over the years.