Business Intelligence and Applications

ASSIGNMENT 2: PIVOT TABLES & TABLEAU

April 1, 2023

Andrea Cardia
Xiana Carrera Alonso
Shradha Maria

Università della Svizzera italiana

Introduction

In this report we aim to describe the conclusions of the analysis of a data set of US cities, with a focus on the patterns and insights gained from it. A more detailed overview of the usage of Excel's Pivot Tables and Tableau is available as a video at this link. Note that because of space limitations, this report only includes the most noteworthy representations amongst all the created ones.

Task 1

Using Pivot Tables (either with a sorting option or a top 1 filter), we can easily identify the requested states: for the most cities, the answer is California, with 580; for the most airports, the answer is Texas, with 2,787.

However, a visual representation of the data with bar charts offers more interesting results regarding the distributions of cities and airports. Though the exact placement of each state may differ between the two, the majority stay in similar positions, which indicates a relationship between the two variables (possibly also related to the states' population and size, which seem to correlate positively with the aforementioned variables).

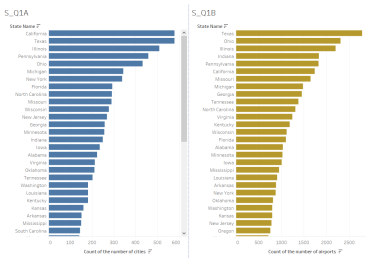

Figure 1: Bar charts of the states sorted by number of cities and airports, respectively.

state_name	Sum of airport_count
Texas	2787
Ohio	2309
Illinois	2203
Indiana	1830
Pennsylvania	1822
California	1742
Missouri	1651
Michigan	1479
Georgia	1456
Tennessee	1381
North Carolina	1316
Virginia	1242

state_name	Count of city
California	580
Texas	579
Illinois	510
Pennsylvania	459
Ohio	432
Michigan	343
New York	339
Florida	294
North Carolina	292
Missouri	290
Wisconsin	278
New Jersey	269

state_name with most airport	Sum of airport_count
Texas	2787

state_name with most cities	Count of city
California	580

Figure 2: Two pivot tables with a sorting option, and two with a top 1 filter, in Excel.

Task 2

By grouping the cities by *county_name* in Pivot Tables, we found that those with the highest population are Queens (18,972,871), Los Angeles (17,602,724) and Cook (10,751,312), whereas those with the highest number of cities are Jefferson (116), Washington (98) and Cook (91).

The number of coincidences between the corresponding top 10s is fairly low, with only 2 counties present in both: Los Angeles and Cook. Even if we extend the limit to 20, few other co-occurrences appear (Jefferson, Wayne, Suffolk and Orange).

Regarding the ranking distribution, towns (3) are the most prominent, which is an intuitive result, since they are much more common than heavily populated urban complexes and our dataset has few instances of villages and communities, with 52 and 1, respectively. We also decided to explore the distribution of population among rankings, finding that the metropolis (1) and mid-size cities (2) are the most prominent, which is also reasonable, as they tend to concentrate the majority of the population.

Finally, we gave a different representation of the rankings for the top 3 counties by population and number of cities with a violin plot. Here, the rankings can be interpreted as a continuous variable, with its peaks emphasized through the use of a kernel density estimation.

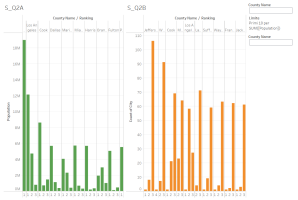

Figure 3: Grouped bar plots in Tableau.

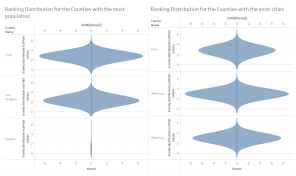

Figure 4: Violin plot in Tableau.

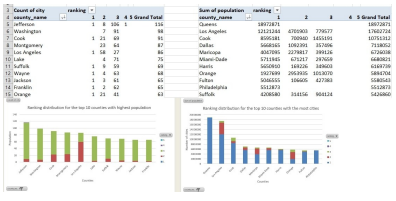

Figure 5: Pivot Tables and stacked charts in Excel.

Task 3

For this question, those cities with mean income 0 or median income 0 were discarded, a situation we interpreted as a "mising value". It is also important to note that there are several cities in different states with

the same name. Therefore, it is necessary to drill down on *state_name* to uniquely denote each one.

Using the absolute value of the fairness score in a Pivot Table, we could observe that the three fairest cities are Adair, in Oklhahoma (with fairness -7.35); Gladwin, in Michigan (-38.58); and Woodlake, in California (59.71).

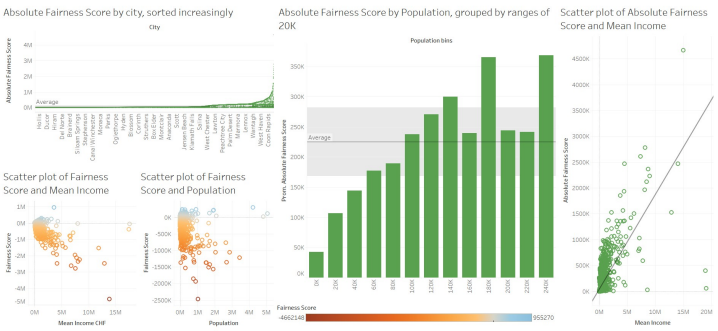

Figure 6: Dashboard for Task 3.

When ordering the cities by absolute fairness score (top left chart in the dashboard of Tableau), some of them really stand out, but most have a low value, resulting in an average of 74,423. This can be attributed to the fact that there is much more data collected about small, modest cities than big, rich ones.

At first, the scatter plots between fairness score and mean income and population, respectively, did not provide much clarity. However, when filtering out the outliers (i.e., the cities with very high values for one or more of those three variables), some traces of linearity started to appear, supported by trend lines. This became clear once we grouped the population and mean income in bins of fixed size, relegating all the outliers with high values to a single bin. Using bar charts and averaging the values for each bin, we could finally affirm with certainty that cities with more population or with more mean income tend to be less fair, that is, have a bigger difference between mean and median income.

Task 4

Since the geographical representation capabilities of Excel were somewhat dissatisfying, for this question we mainly used Tableau (much more precise than Excel's map - see task 8). By plotting the top 100 cities by population, it can be observed that the most salient ones are situated on the west and east coasts, or near other masses of water (lakes, rivers...), with only a handful of exceptions in the interior of the country. It is also worth noting that many cities are close enough to facilitate transport, commerce, etc., but with enough space in between to not directly compete for resources.

We added details such as the state names and city names as details to the plot. The population is indicated both as a detail and through the size of the points.

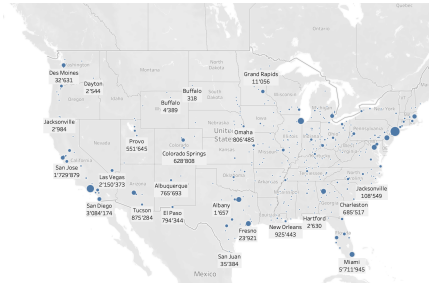

Figure 7: Map plotted using latitude and longitude data.

Task 5

Using the standard `CORREL` function in Excel, we computed the correlation values for the relevant pairs of variables (all converted to metres and CHF for easier interpretation, which does not affect the results). *Population* and *Available Land* (m^2) have a correlation of -0.0045; *Median Income* (CHF) and *Average Airport Score*, of 0.41357; and *Available Water* (m^2) and *Average Airport Elevation* (m), of 0.0669. The first and third values are really close to 0, indicating that no linear relationship exists between the variables. The second one has a higher (and positive/direct) value, but not high enough to be considered specially significant.

Figure 8: Scatter plots created in Tableau.

The scatter plots created justify and support the mentioned results. The points for the first pair are predominantly situated on the X or on the Y axis, forming an almost horizontal and almost vertical line, both of which are indicative of no correlation. The points of the third pair are really close to a horizontal line on the X axis, also showing no correlation. The second pair again presents a strong verticality near the Y axis, but the points are sufficiently scattered around so that they suggest some direct linearity (i.e., more airport score is related to more median income). However, the strength of the vertical segment is remarkable enough, compared with the pattern formed by the other points, to justify discarding correlation in this case too.

Task 6

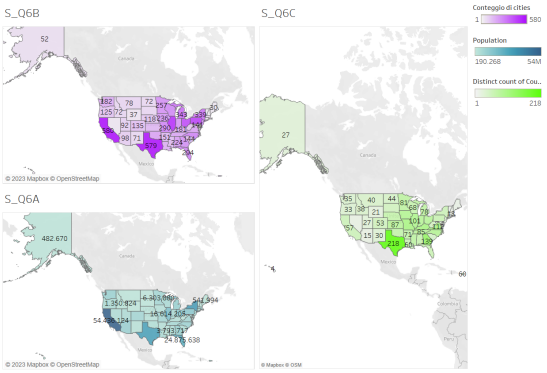

Figure 9: Dashboard for Task 6 in Tableau.

The three most populous states are California (54,436,124 people), New York (31,318,308 people), and Texas (29,620,776 people). Respectively, they have 57, 57 and 218 counties, and 580, 339 and 579 cities. Note that since some cities share the same county, they had to be aggregated using `COUNT DISTINCT`.

A map allows us to see that there are striking differences in the population of states. Those mentioned before, and some others like Florida, stand above all the rest, whereas most of the states to the interior of the country are far less populated. The number of cities follows a similar distribution, although the difference between coastline and interior is not so remarkable. With respect to the number of counties, Texas is far above all others, with almost a hundred counties of difference with the second state in that category. The vast majority of states have less than a hundred counties, or barely surpass the mark.

Task 7

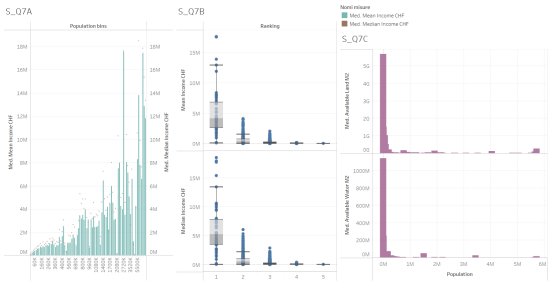

Figure 10: Most salient visualizations for Task 7.

Due to the results obtained in exercise 3, we hypothesized that the median income tends to be higher than the mean income. Indeed, by comparing population and these two variables, both in the leftmost figure of this exercise (in which the points, that represent the median, are usually higher than the bars, that correspond to the mean, when using symmetric axis) and in the scatter plot done in exercise 3 (in which the majority of cities showed a negative score), we can infer that the median tends to always be higher.

This led to a study between the correlation of income and population. Mean, median and standard deviation of income have correlations of 0.77, 0.74 and 0.78 with population, a really significant indicative of a direct relationship between them. We can expect that the higher the population, the higher the income, but also the higher the variability of it among the inhabitants. This is also reflected when drilling down by ranking (see the boxplot graph).

We also hypothesized a relationship between the available land and water. Indeed, their distributions show similar results, suggesting that cities with a large territory also tend to have large reserves of water. Other studies, such as the relationship between airport score and its elevation, gave no indicative of relation and were discarded.

Task 8

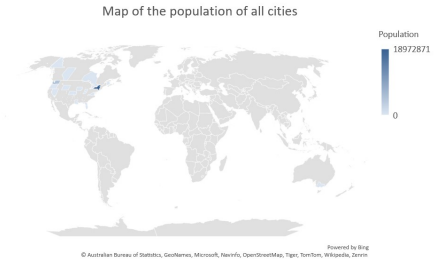

Figure 11: Map plot made in Excel for question 6, showing a low accuracy in the cities placement.

In general, Tableau offers more visualization options, with a higher degree of interactivity and customization. We found it easier to do tasks of preliminary analysis of variables in it, since creating a graph needs less steps, speeding up the process of gaining a general insight of the results. Excel's Pivot Tables, on the other side, have more tools related to the manipulation of the variables, making them more suitable for tasks of aggregation, drilling down, rolling up, filtering, sorting, performing standard calculations (differences, correlations...), etc.

Because of these reasons, each tool has its limitations. For example, the maps that we created in Excel were quite inaccurate, since it has no option to use latitude and longitude variables but instead it plotted the information based on the names of cities, states, etc. Additionally, many plotting features are not available for Pivot Tables, so data needs to be copied outside of them. Conversely, Tableau could not give us the exact values of correlations between variables, an almost trivial task in Excel, and sometimes was not as intuitive to use.

Therefore, we conclude that both tools are useful and powerful for data analysis, covering each other's weaknesses and offering a wide variety of interpretation possibilities for the analysts that choose to combine them.