Report

Assumptions:

1. Medtronic Russia Office is located in Moscow. My distance assumption is based on Moscow is the starting point and distance of each city is calculated from Moscow. Units in Metric(Kms)
2. Measurement is done in Metric as Russia used the metric system
3. Sales Rep data is limited. Here the assumption is that the sales reps assigned to any cluster reside in one of the cities in that cluster to minimize travel to and from Moscow.
4. Formatted the dataset. Removed “null” entries from city column. The city/hospital “names” were first translated to English and then formatted with standard measures like removing “null” values, whitespaces, tab breaks, leading and trailing characters. There is no other dataset to verify against if these names are correct or not. The assumption is that the dataset after cleaning is accurate to the best of the given circumstances.
5. Few of the cities are in the disputed location (Crimean annexation by Russia). Simferopol/Sevestopol have been added as Russian territories. The assumption is that travel between these and other cities are not restricted.
6. Missing data in the PCI sales columns have been updated with 0. There are many columns with missing or null values. I believe using a mean/median value for these missing values would skew up towards a higher side of estimations
7. Only dataset. No other dataset available to verify/cross tabulate with. I am assuming that the source of this dataset is accurate at the onset.
8. Starting sales rep calculation with value 20.
9. At least 1 sales rep per cluster. Cannot use partial number while defining a number of humans.
10. employee cost/sales rep location/profit data etc not available

Steps:

1. Cleaned the dataset with English city names from Russian. Used Google sheets/excel for translation.
2. Formatted/Removed “null” entries from city column. The city/hospital “names” were first translated to English and then formatted with standard measures like removing “null” values, whitespaces, tab breaks, leading and trailing characters.
3. Grouped dataset by Category/city/hospital name.
4. Mapped the Cities in Tableau. Added a few cities as Russian territories. (Geographic coordinates)
5. Calculated the distance (in kms) from Moscow for each city.
6. Using Alteryx, calculated K-Clustering based on PCI sales data (forecast and fact) and Distance from Moscow as the predictor variables in Alteryx. The result indicated 9 as the optimal number of clusters. The data for the predictor variables were standardized using the z-score to minimize the effects of outliers. The clustering method used was K-Means. The bootstrap sample size for each iteration was 50.
7. Once the clusters were defined, calculated the sales per cluster, per city, counting a number of hospitals. Then the percentage of hospitals per clusters was calculated. Along with the percentage of fact over the forecast for 2015 was calculated. This measure gives an estimate where sales were lower/higher compared to the forecasted sales data.
8. Finally, the allocation formula I used was, with 50% weightage each for percent of hospitals per cluster and percent of fact/forecasted sales per cluster, multiplied by the “assumed” sales rep number of 20.
9. The result was rounded of to the nearest upper integer value, as we cannot use partial numbers to define human count.
10. If this same result value is floored (i.e chose the lower integer value) for rounding, the minimum number we get using this value is 16. So, if employee cost (overhead cost) is a top factor, we could use a minimum of 16 reps for these locations
11. The optimal number of sales reps to cover the cities(clusters) is 25.

Suggestions for further analysis

1. Gather as much data as possible. Gather data regarding sales representatives i.e location, sales targets forecasted/ completed, etc

2. Gather more data about products/hospitals and what factors drive sales. Every region is unique and needs an understanding of local markets for better market share.

3. Understand relationships between doctors/patients/sales representative.

4. A better understanding of company policies and management decision making such as priorities/goals and deliverables. For example, is the priority is market share capture or is it cost optimization to survive the competition. etc

PS: This list is not limited. With time and better understanding, the iterative nature of data analysis can improve the bottom line.