**G53IVP Coursework 1.1 Analysing Data with Tableau System**

**Name : Yap Cheng Yee**

**Student ID : 016759**

**1.0 Hypothesis**

**Hypothesis/Tasks for 1st Dataset:**

1. Identify the country with the highest prevalence of tuberculosis
2. Identify the region with the highest prevalence of tuberculosis
3. Higher number of populations indicate higher number of deaths related to tuberculosis
4. Identify the country with the highest number of deaths from tuberculosis without HIV
5. Identify the country with the highest number of deaths from tuberculosis with HIV
6. Whether tuberculosis or HIV is the primary reason of increase of death?
7. Does the number of cases or incidence of tuberculosis increase over the years?
8. What are methods used to derive tuberculosis and HIV estimation?

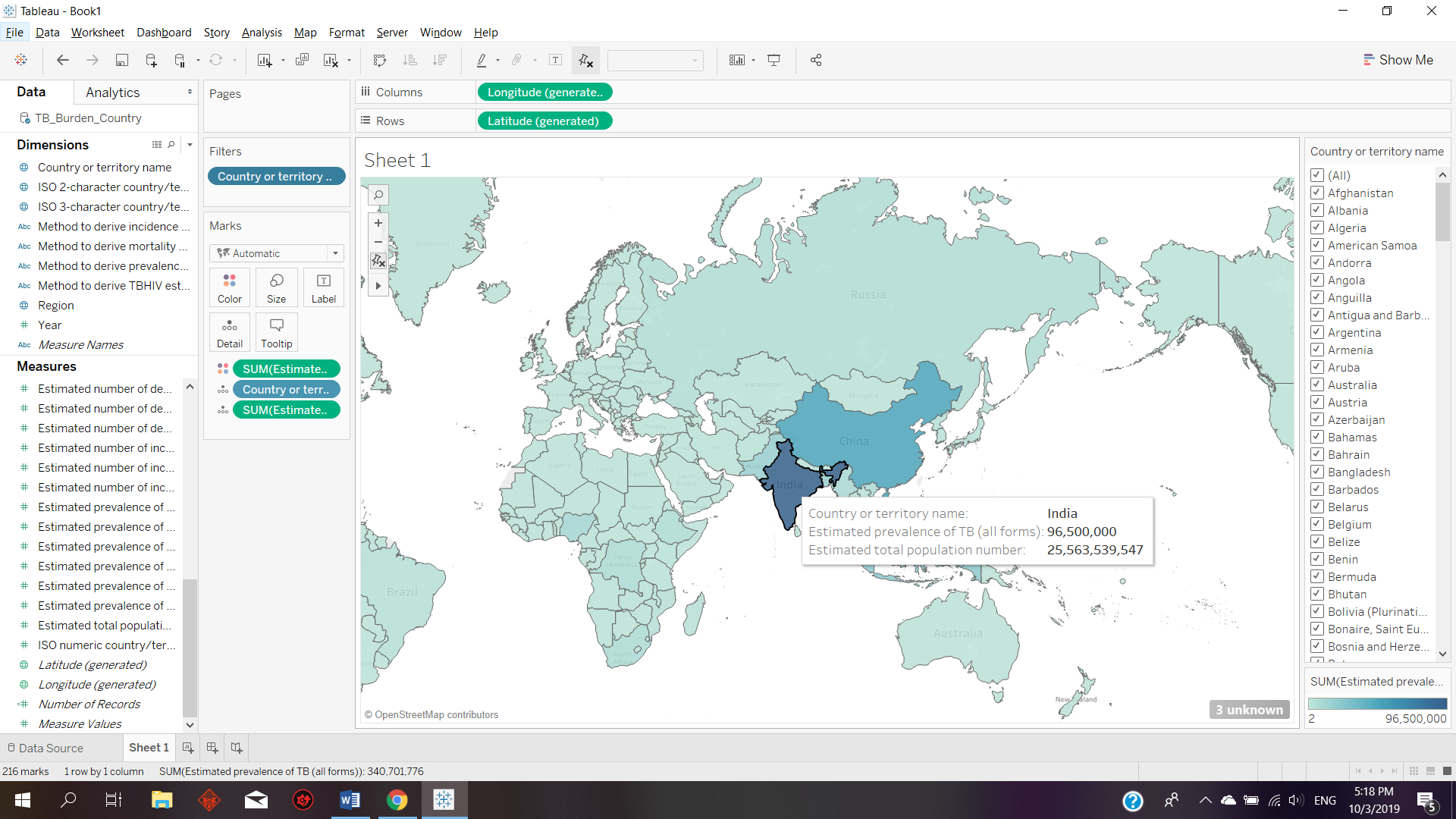
**Hypothesis/Tasks for 2nd Dataset:**

1. Which room type contains the highest rating?
2. Which neighbourhood contains the most review?
3. What room type is most commonly found in specific neighbourhood?
4. Apartment is the highest demand of AirBnb room type as it is simple and not too pricey

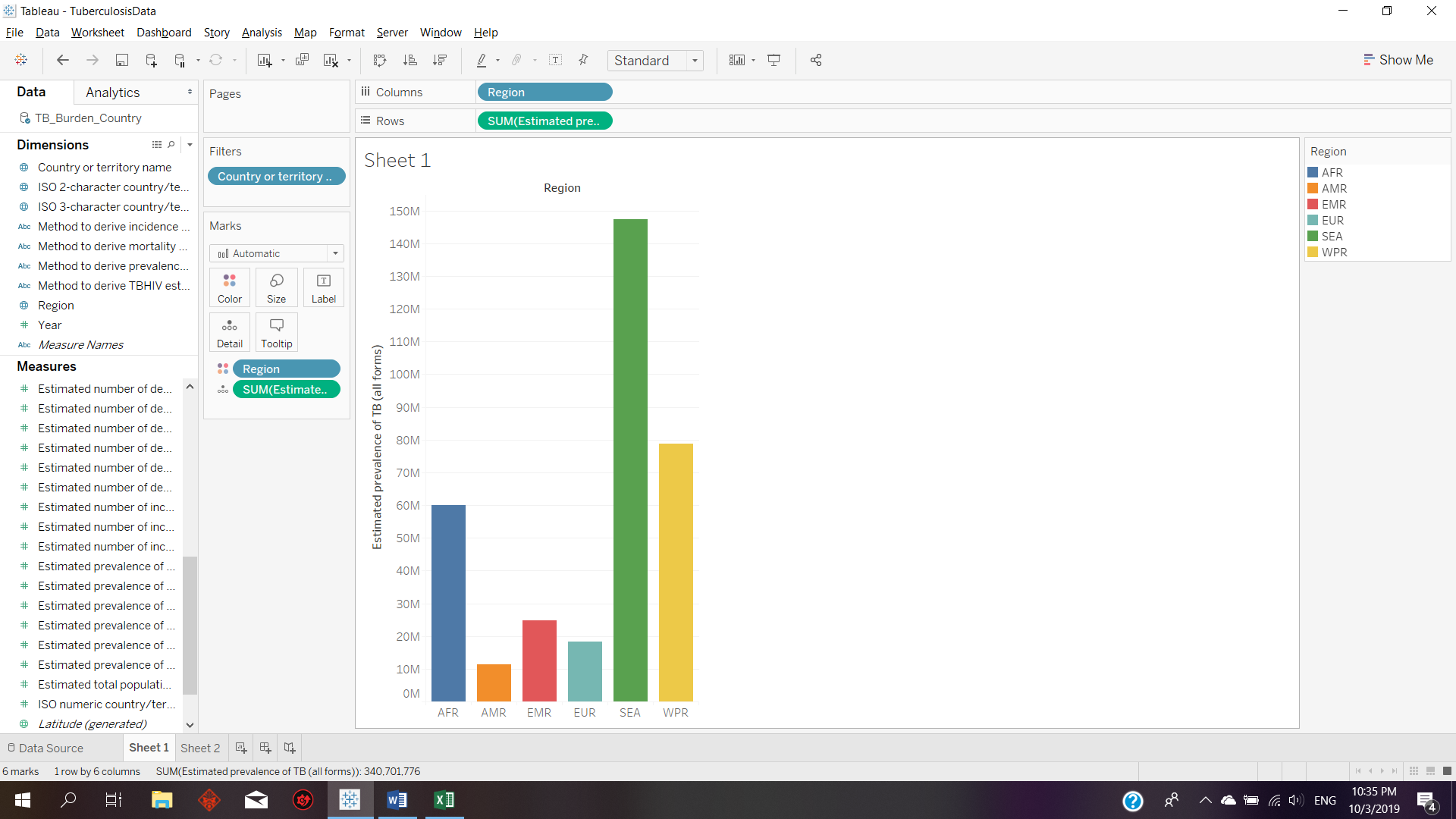
**2.0 Summary of Exploration Process / Data Analysis**

**a) Tuberculosis Dataset**

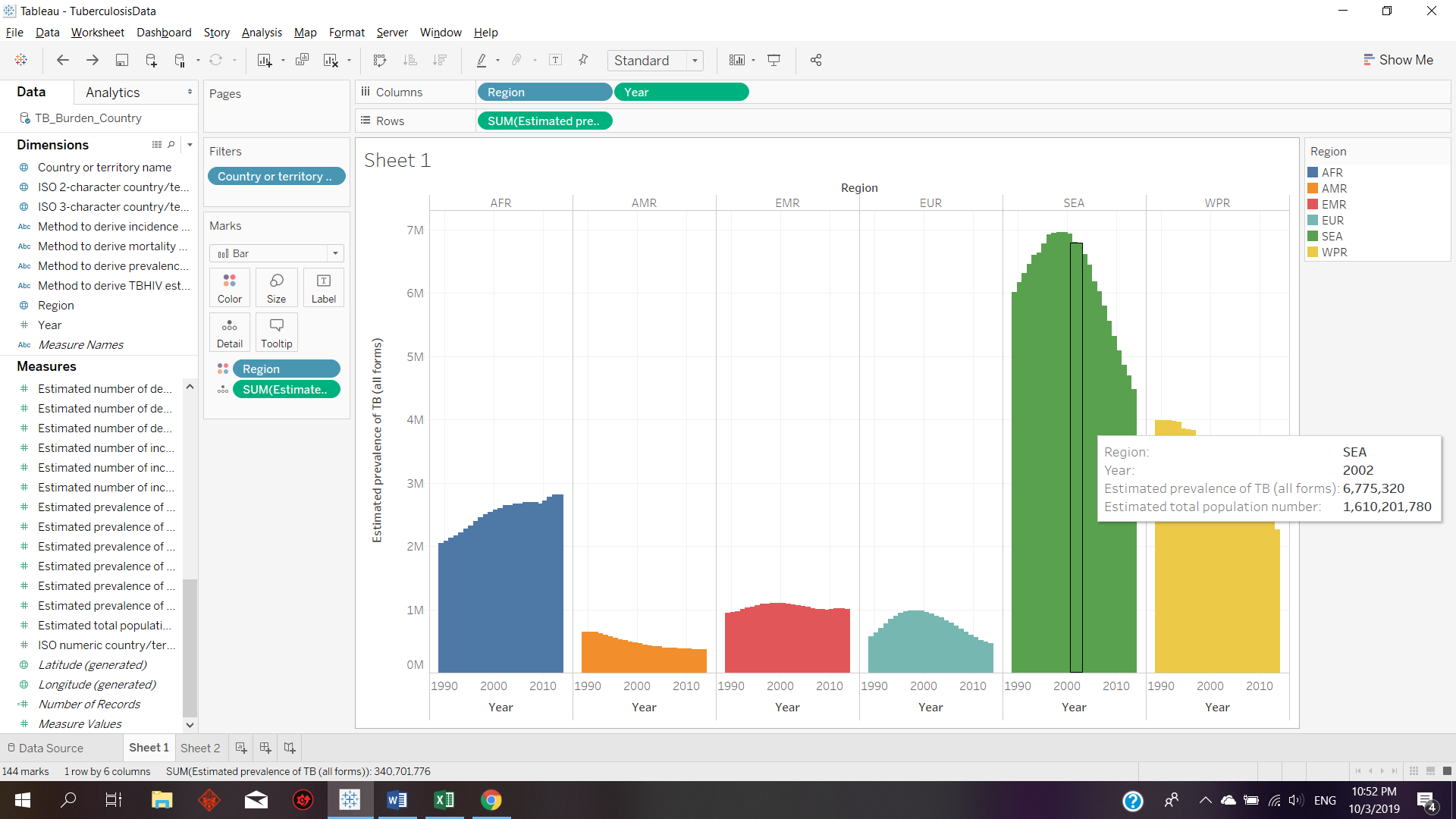
With the use of Tableau, i am able to instantly determine countries that have the highest prevalence of tuberculosis, which would be India. Not far away from India would be China which comes in second in having the highest prevelance of tuberculosis. With the generated map and using colour hues to represent these data, I am able to instantly catch out and find out countries which has high prevalence of tuberculosis based on the hues of the colour.Not only that, if I hover towards the country, I am able to see the exact number of people who has tuberculosis and compared it with the total number of population of the specific country.



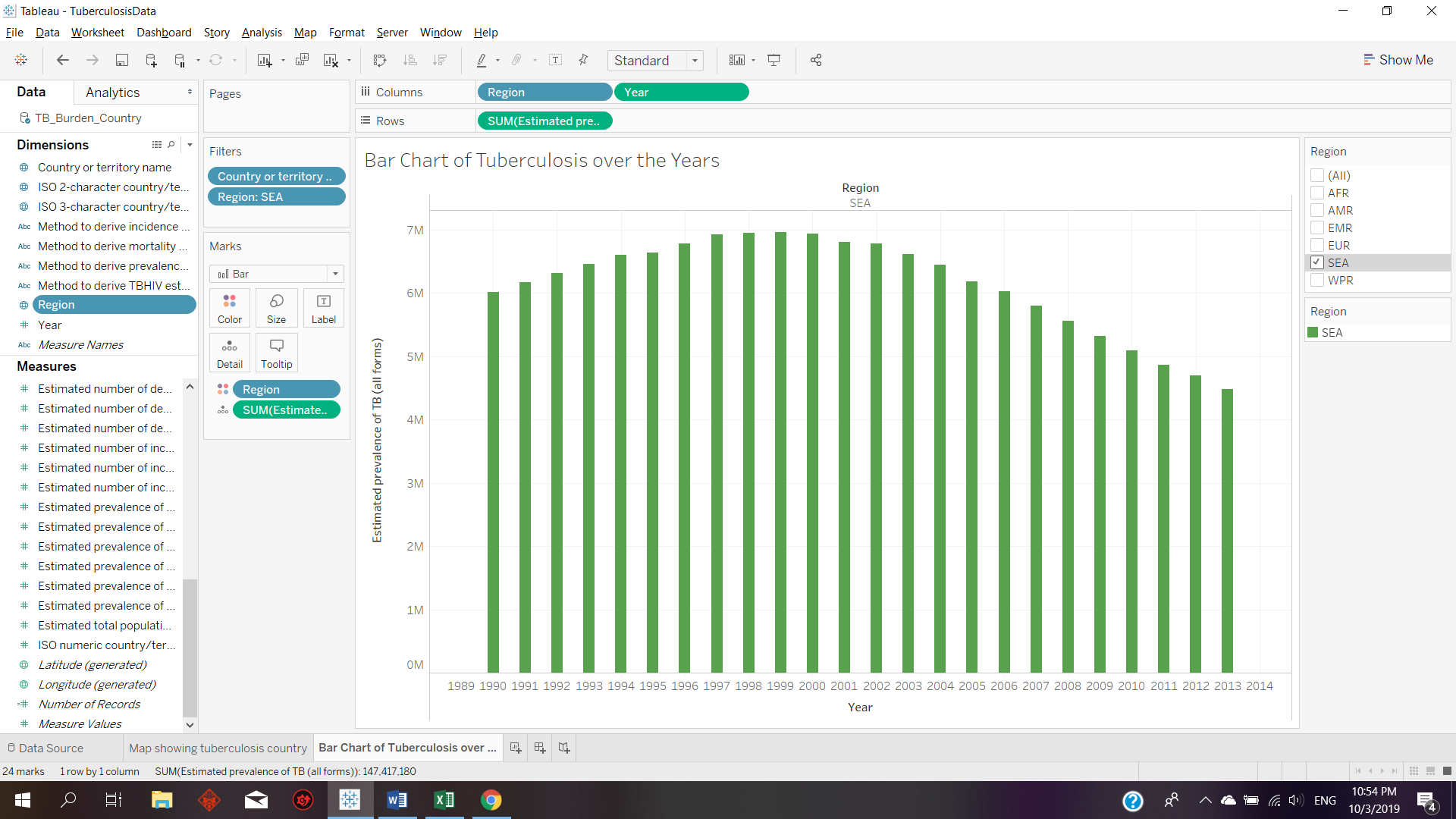
In order to show which region has the highest prevalence of tuberculosis, a bar chart was generated using Tableau where each different colours of bar represent their respective regions. Doing so, we would be able to see that SEA and WPR has the highest prevalence of tuberculosis followed by AFR. This would further support the coloured hue map shown above as India is located in SEA whereas China is located in WPR.



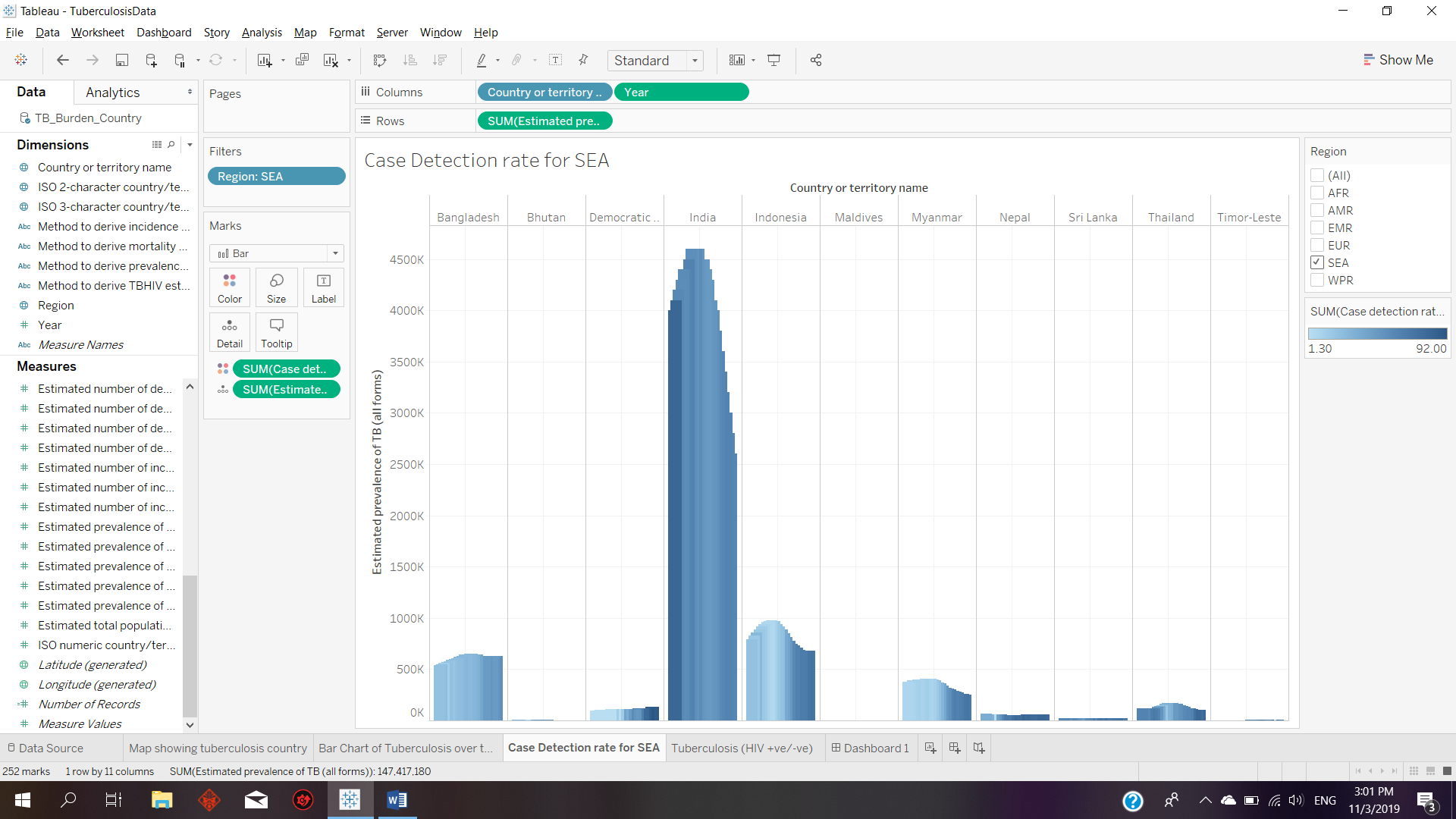
In order to determine whether my hypothesis whereby a does the rise of tuberculosis is affected by the increase of total population, a bar chart is generated and shown that the total prevalence of tuberculosis in most regions are decreasing over the years despite the increased growth of human population. An unexpected finding from the bar chart is that only the African region has a steady increase of prevalence of tuberculosis over the years.



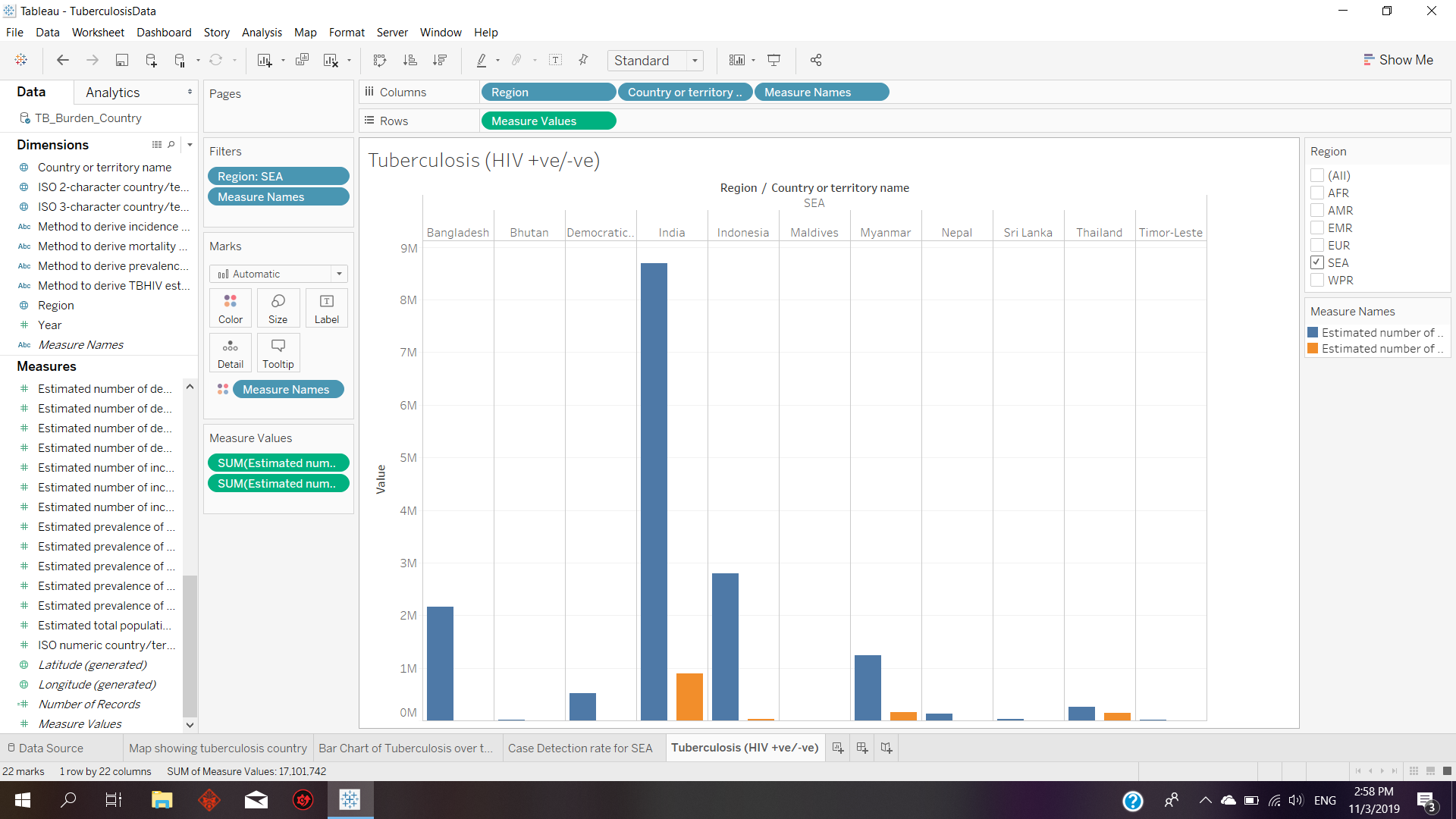
To get a clearer view during the comparison between the prevalence of tuberculosis between each year, Tableau also provides a filter function so that we could select our targeted or interested region.



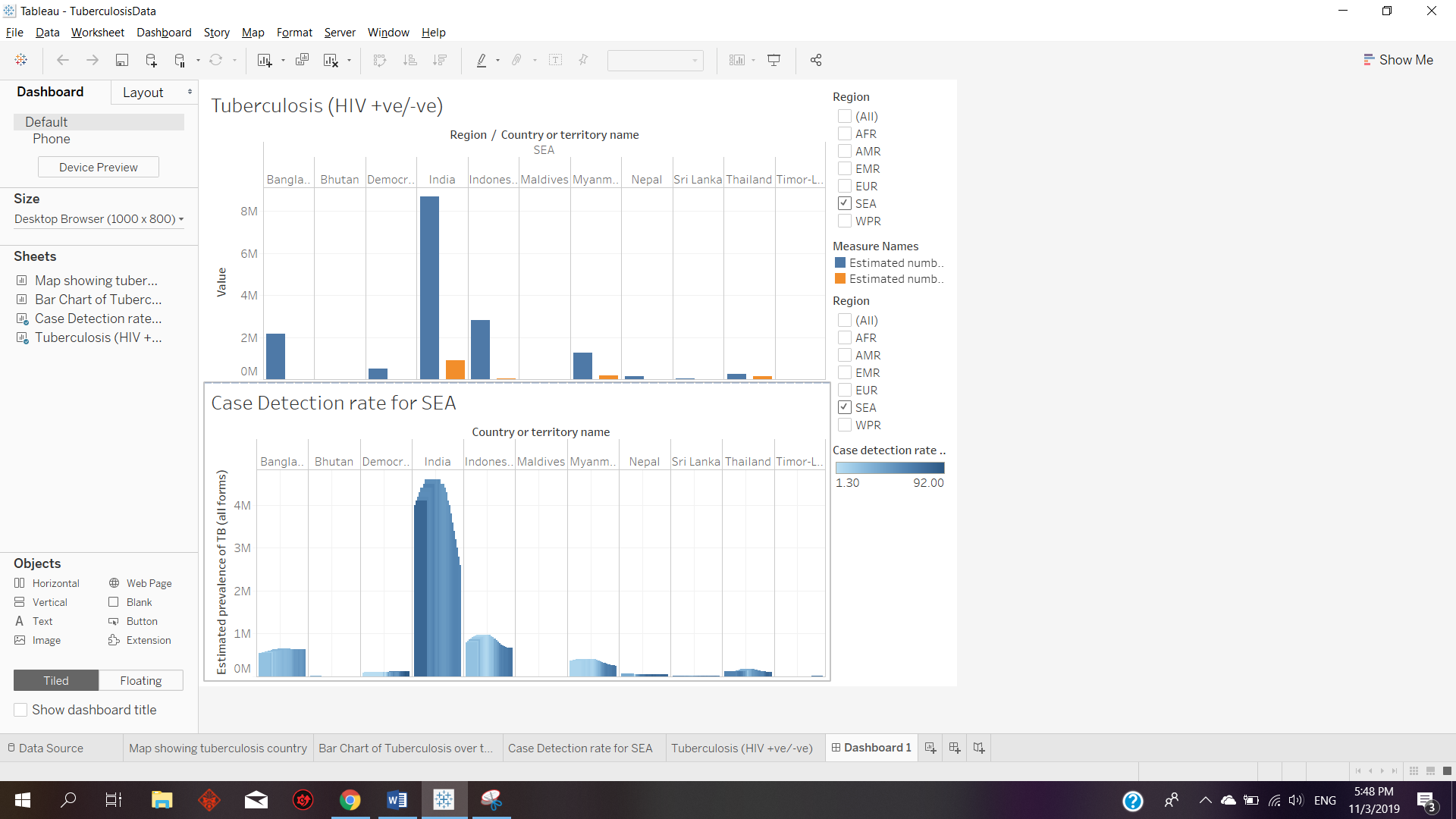
After filtering out the rest of the region with only SEA selected, using Tableau I am able to duplicate the worksheet so that I could continue changing the wanted data’s without having to repeating the same steps. In order to find out whether certain countries contain higher case detection rate due to fact that they could possibly be more advanced, I have added the variable using colours to show the case detection rate over the years where the darker hue would be a higher case detection rate where a lower hue would indicate otherwise.



In order to determine whether most death caused by tuberculosis is tested as HIV positive or negative, a comparison bar chart is created where blue represents death caused by tuberculosis without HIV whereas orange represents death caused by tuberculosis which is HIV positive. With this data, users could determine whether the primary cause of death in the specific country is due to HIV or tuberculosis and to find out the relationship between these two diseases.



Next, to find out the relationship between tuberculosis and its detection rate I have also used Tableau’s dashboard to place two different charts together in a single view in order compare and view their data. This way, I do not have to navigate back and forth in order to compare the data.

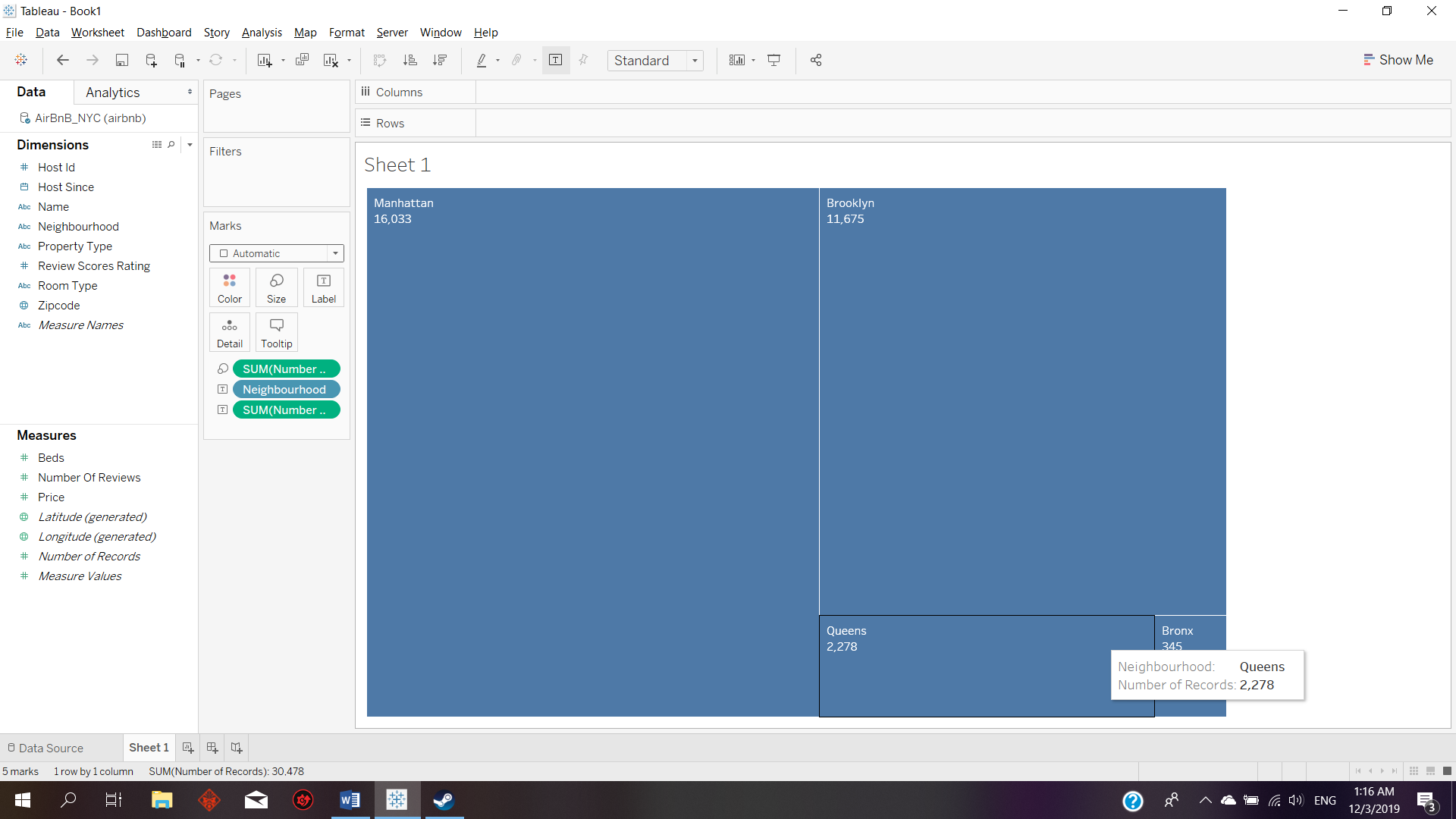


Finally, in order to display all these data to be used for presentation, I have created a Story using Tableau. The story will be explaining the whole idea of the presentation which discuss the estimated prevalence of tuberculosis and towards to country and region which contains the highest rate. From there onwards, using SEA as the region tuberculosis that are tested with HIV positive or negative data are also shown including the case detection rate in a dashboard. Tableau provides the option to enter the caption which could ease the readability of the story for the users. Presentation of the data is achieved by selecting windows and then go to presentation mode.

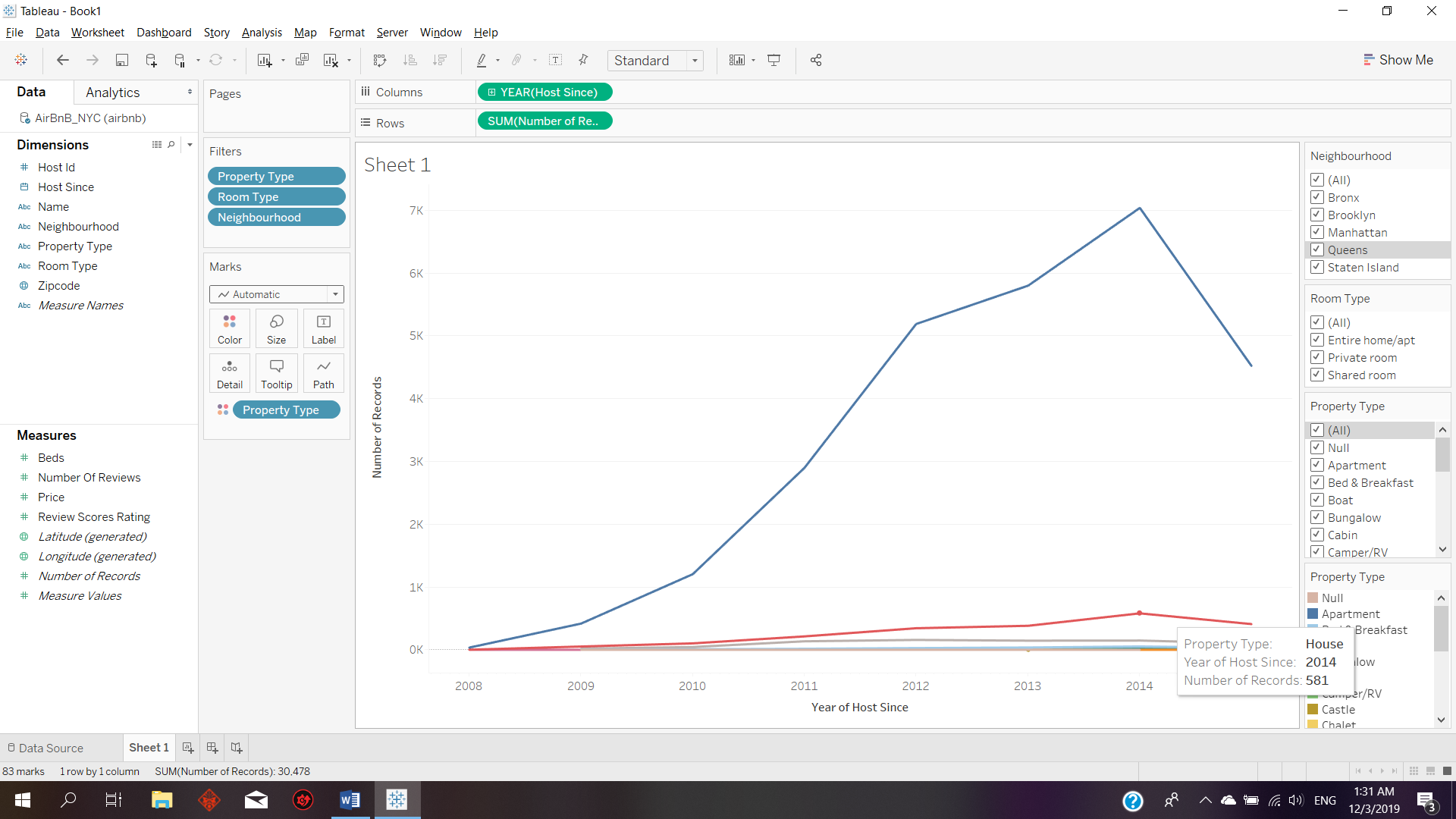


**b) Airbnb Listings in New York City Dataset**

In order to find out which neighbourhood contains the highest record of AirBnb reviews, treemap is used to show users as it is faster and clearer to see. Each size of the block is enough to show the difference between each neighbourhood.



Finally, in order to show the trend and popularity of each property type based on the number of records, a line graph is used to show that over the years even though apartment has a slow decrease from year 2015 onwards, it is still the most used AirBnb property type compared to the rest. This data would be meaningful for tenants who wished to convert their places to AirBnb and they could consider selecting different property types based on the graph.



**3.0Critique of System**

**3.1Advantages of Tableau**

Tableau is a software that can connect to various types of database for example Microsoft Excel, JSON file, or CSV file. This would help users as they do not have to convert their dataset to only a specific file type.

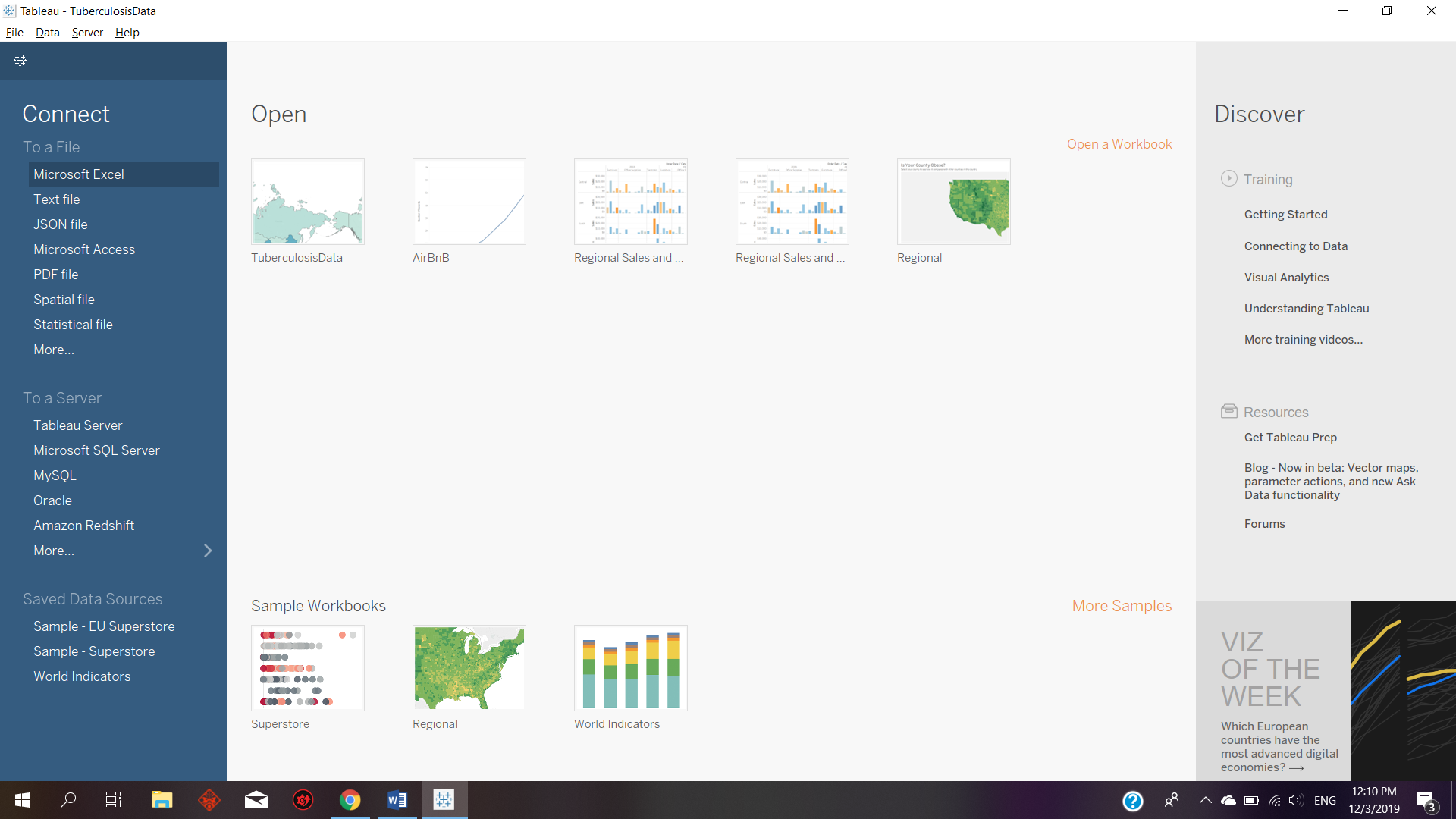
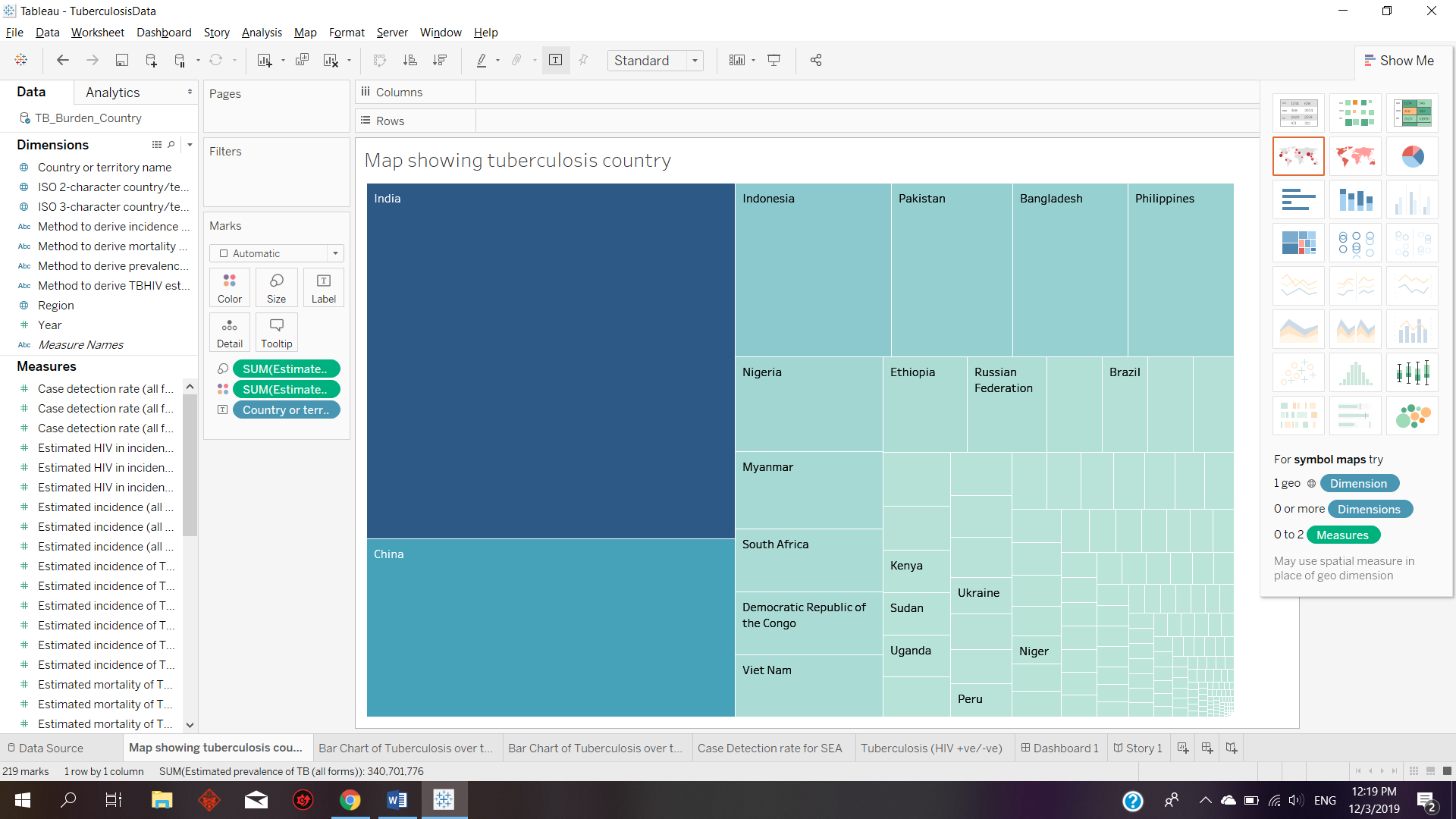


Tableau also allows users to change the dimension or variable name of the dataset in the system. This is a good advantage as if the dataset is not perfect then the variables would be represented by incremented values of alphanumeric which is hard for users to use the variables to design their data visual.



Another advantage of Tableau is that users can change to different visualisation technique in a click away by clicking on Show Me. The same data with variables is maintained and can be represented in different methods for example Prevalence of Tuberculosis in different countries can be visualized using treemaps or maps.



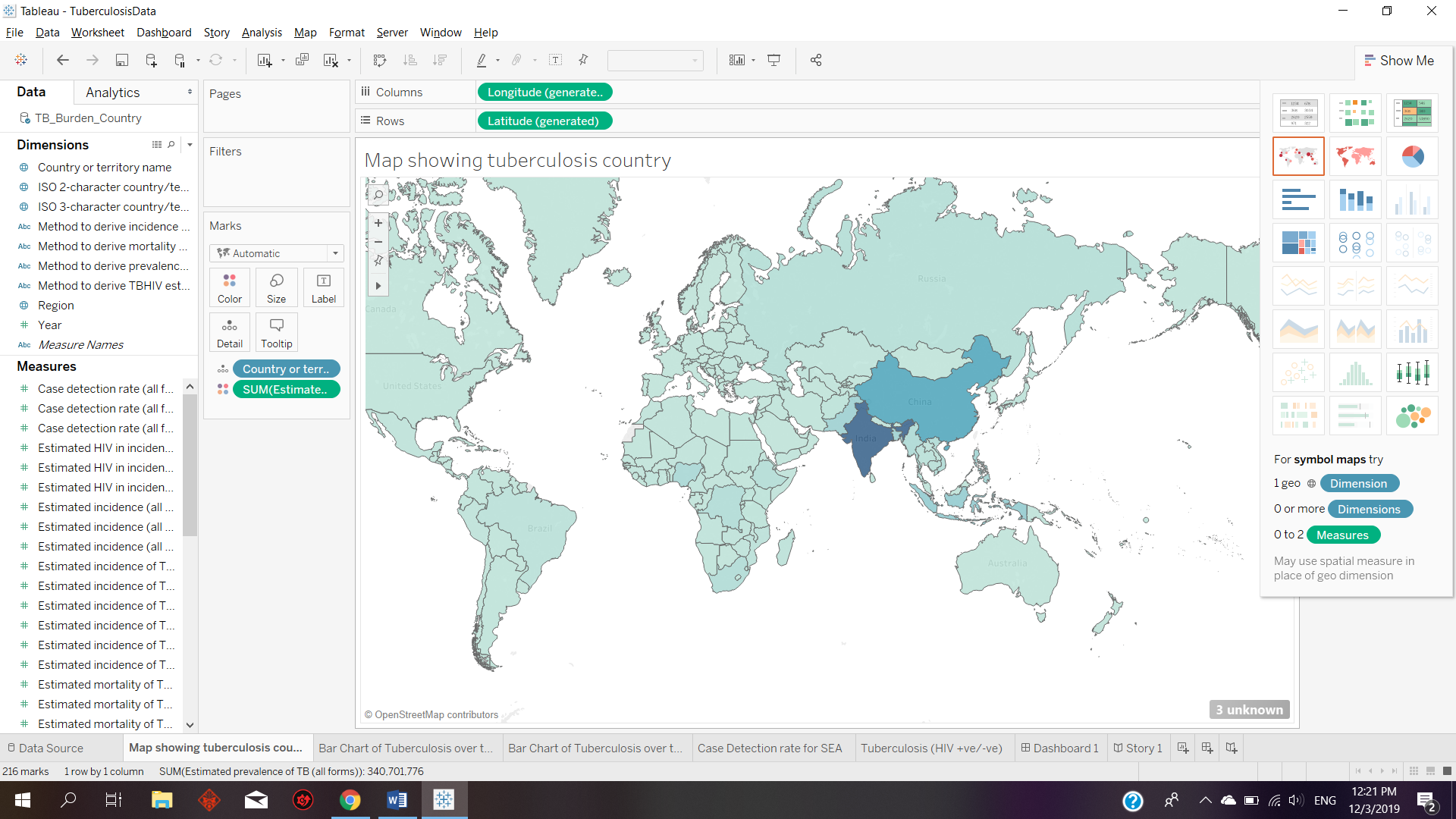
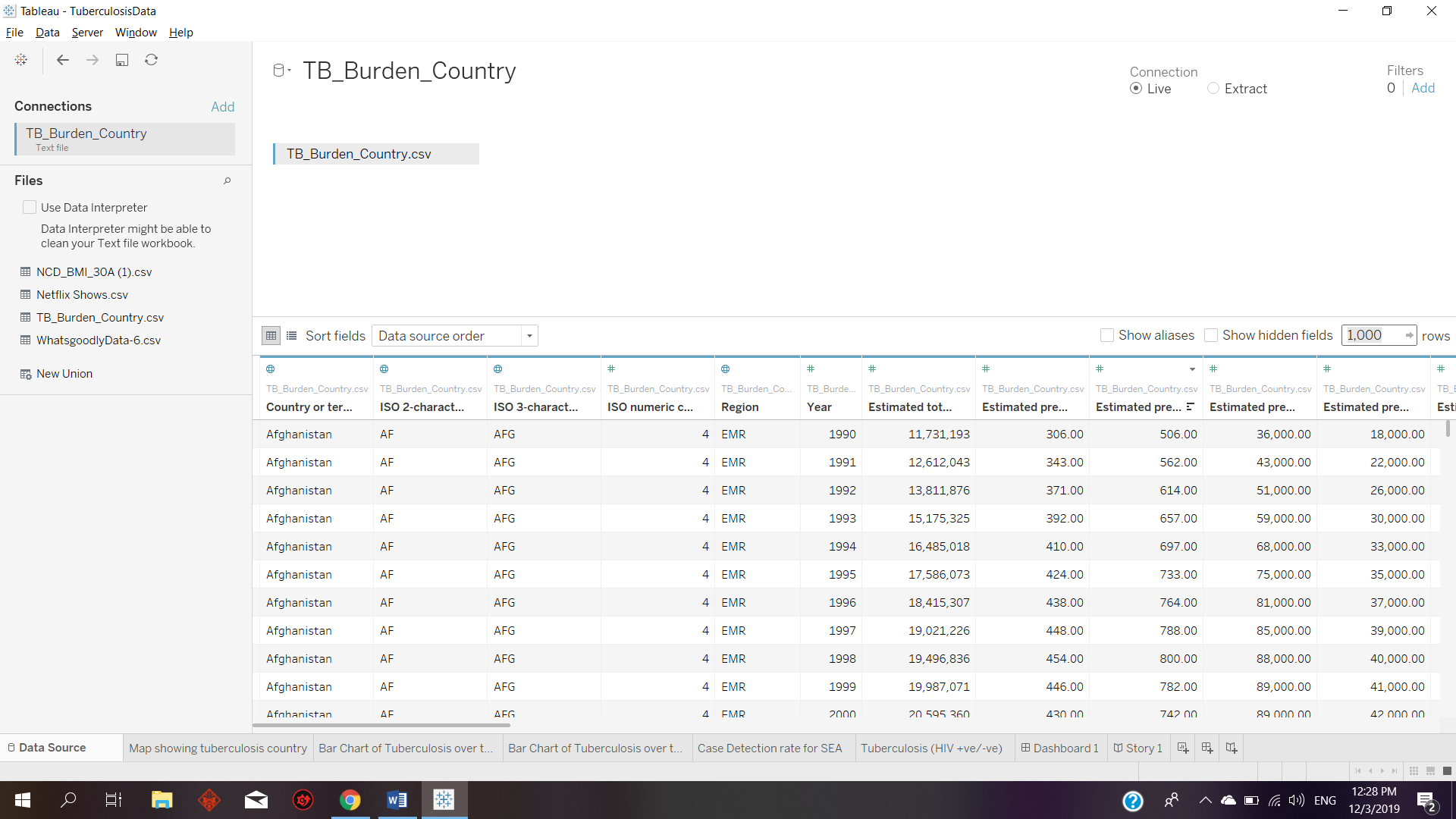
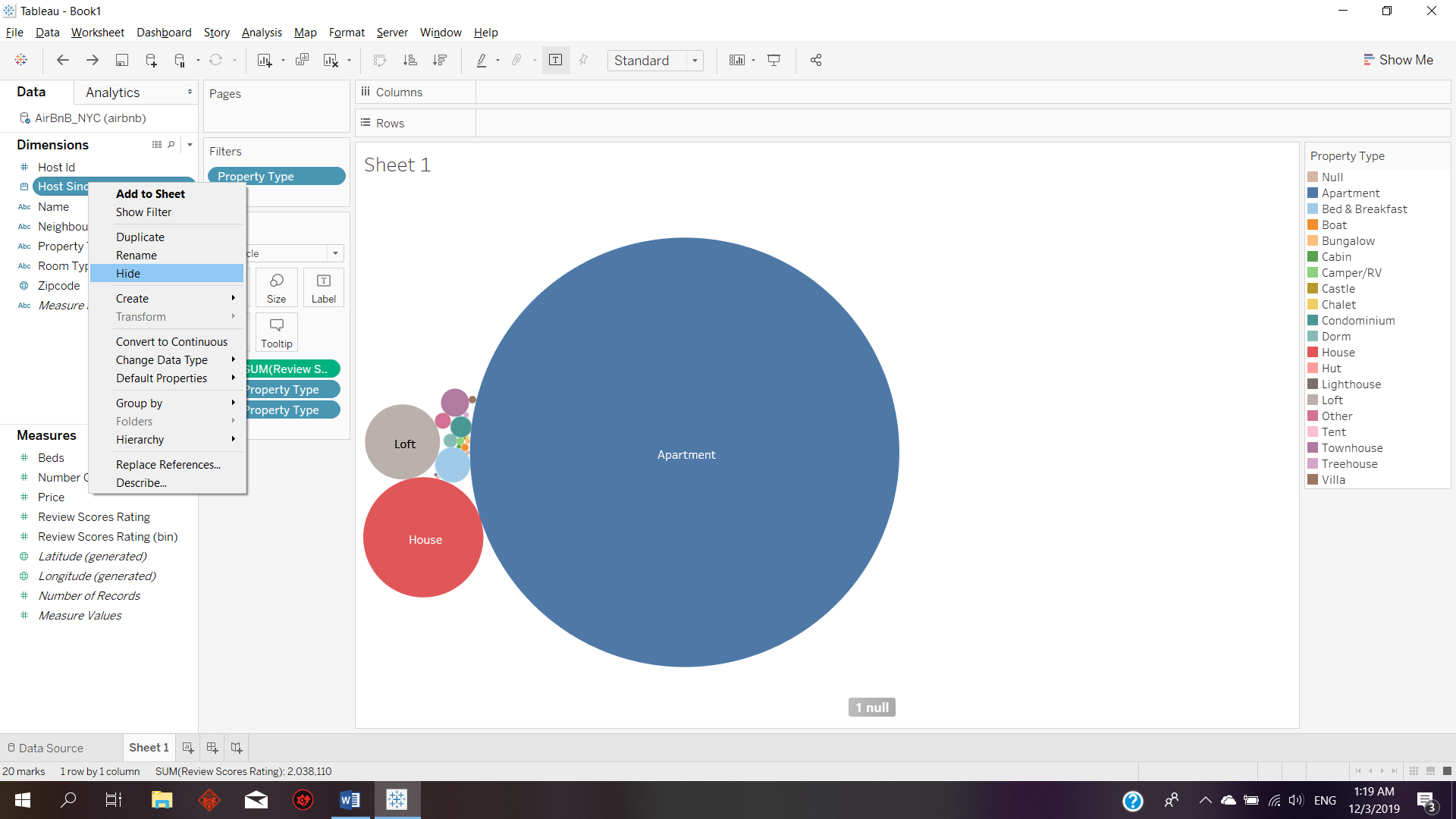


Tableau can also handle large dataset for example the Tuberculosis dataset contains multiple rows and columns.

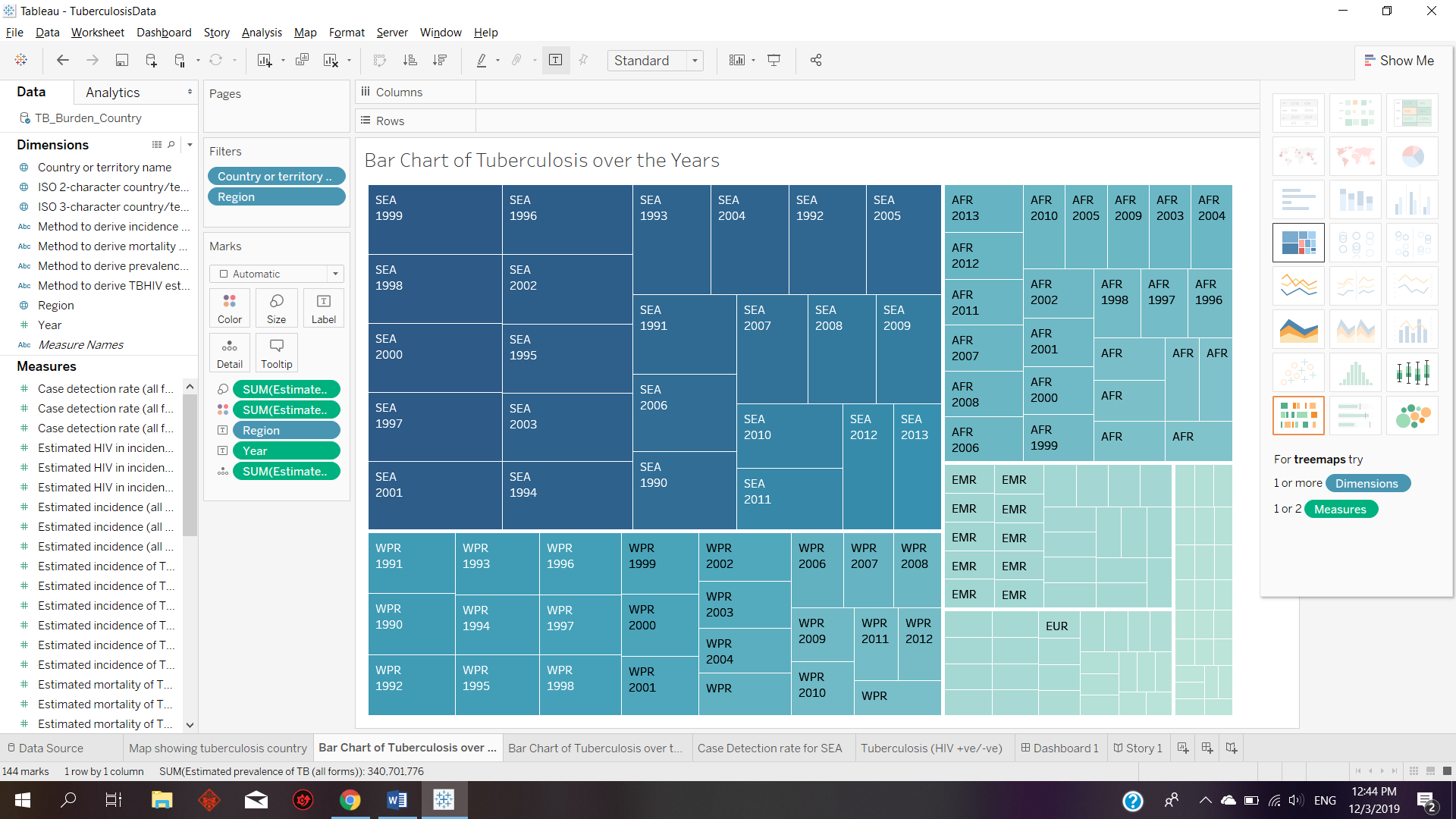


If there are unneeded Measures or Dimensions, Tableau allows the user to hide them. Doing this would make the dimension clearer and lesser values so that it will ease the user when they are searching for their intended variables.



**3.2 Disadvantages**

The disadvantage of Tableau would be the incapability of merging different dataset. Tableau requires user to have a proper dataset before connecting to it only then the system would perform as intended. Another disadvantage is if there are certain fields added for specific data visualization techniques, changing it would revert the additional fields and user might miss out some of the details as Tableau will automatically add relevant field based on the selected data visualization techniques.



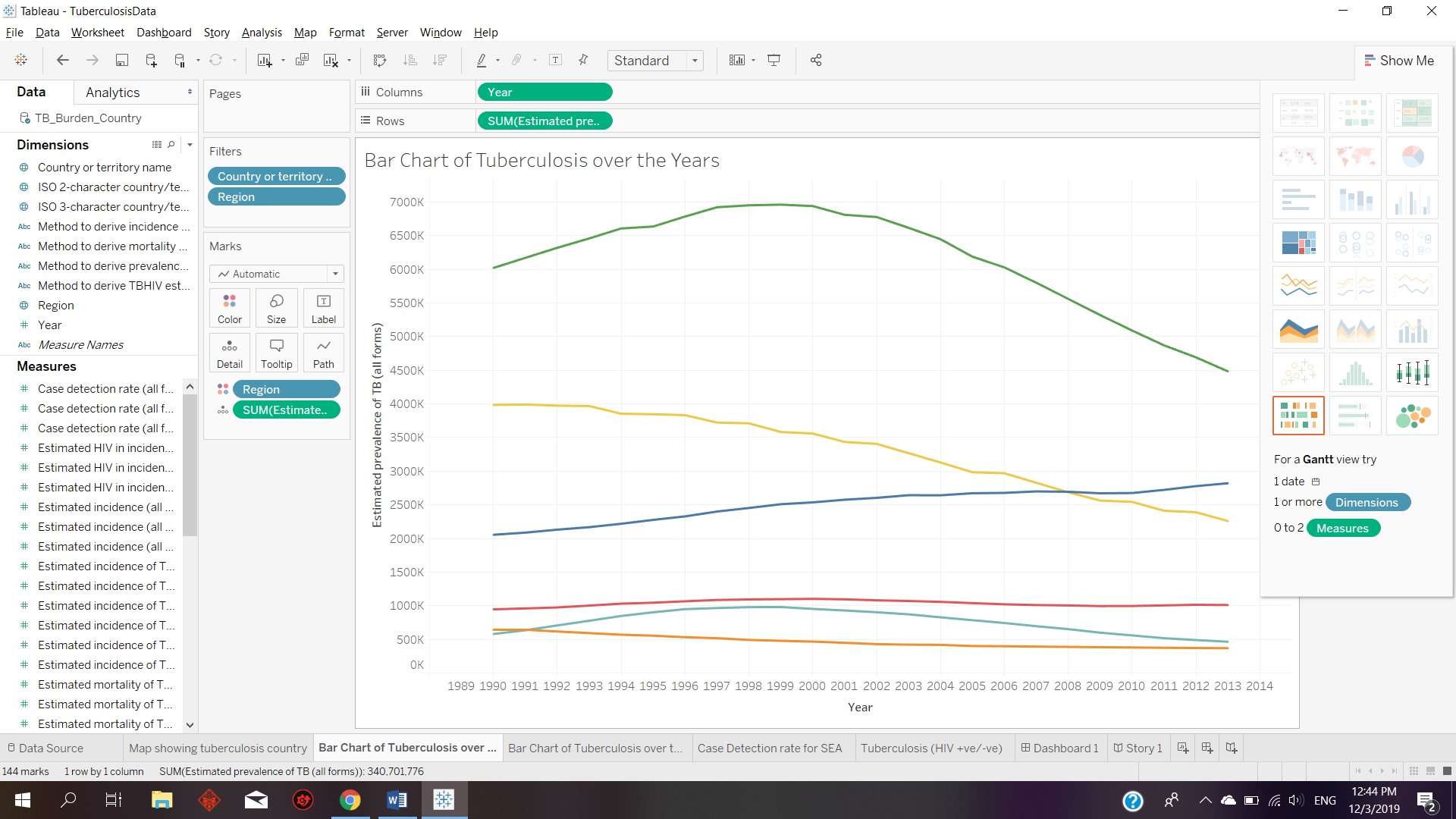
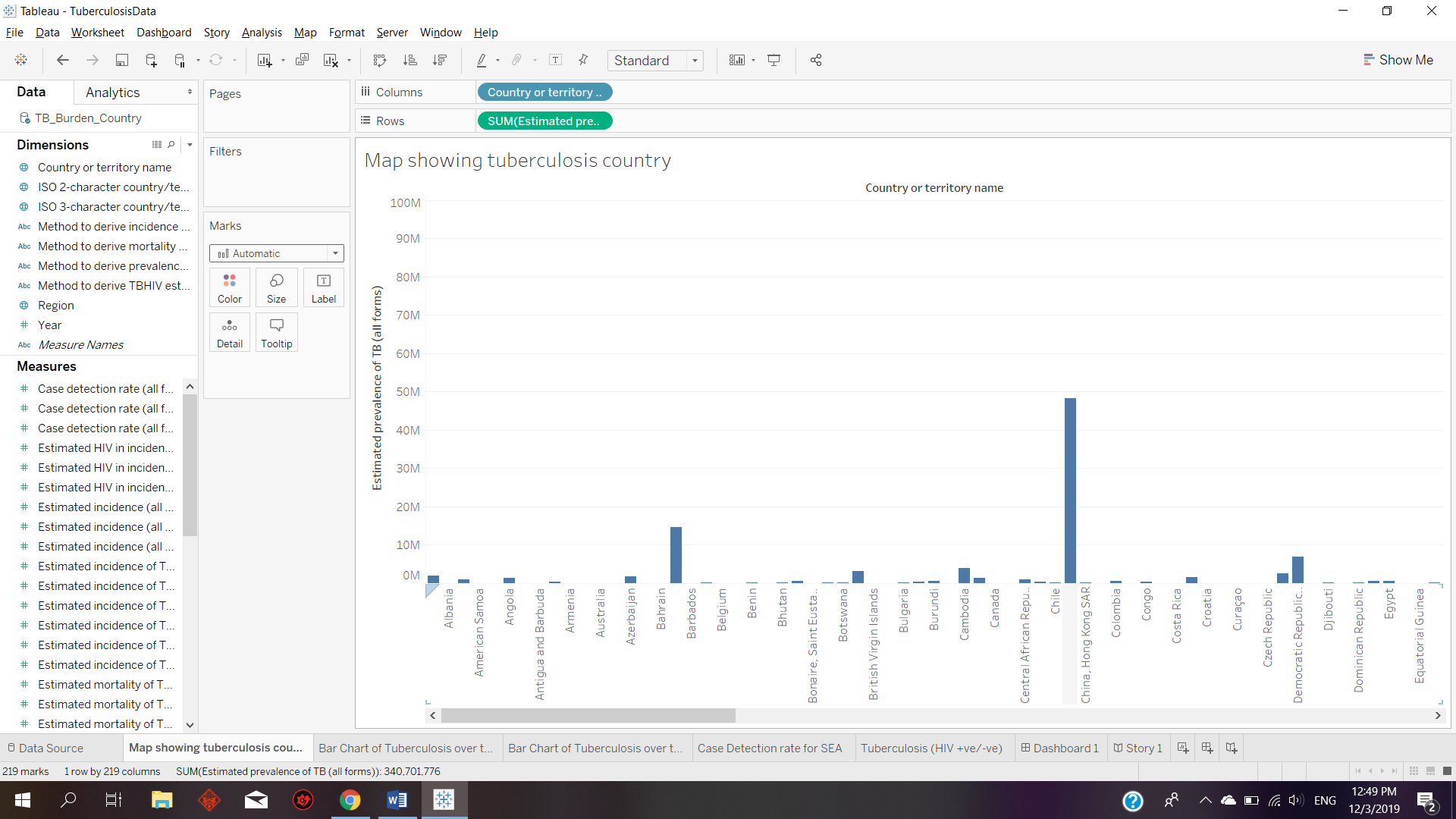
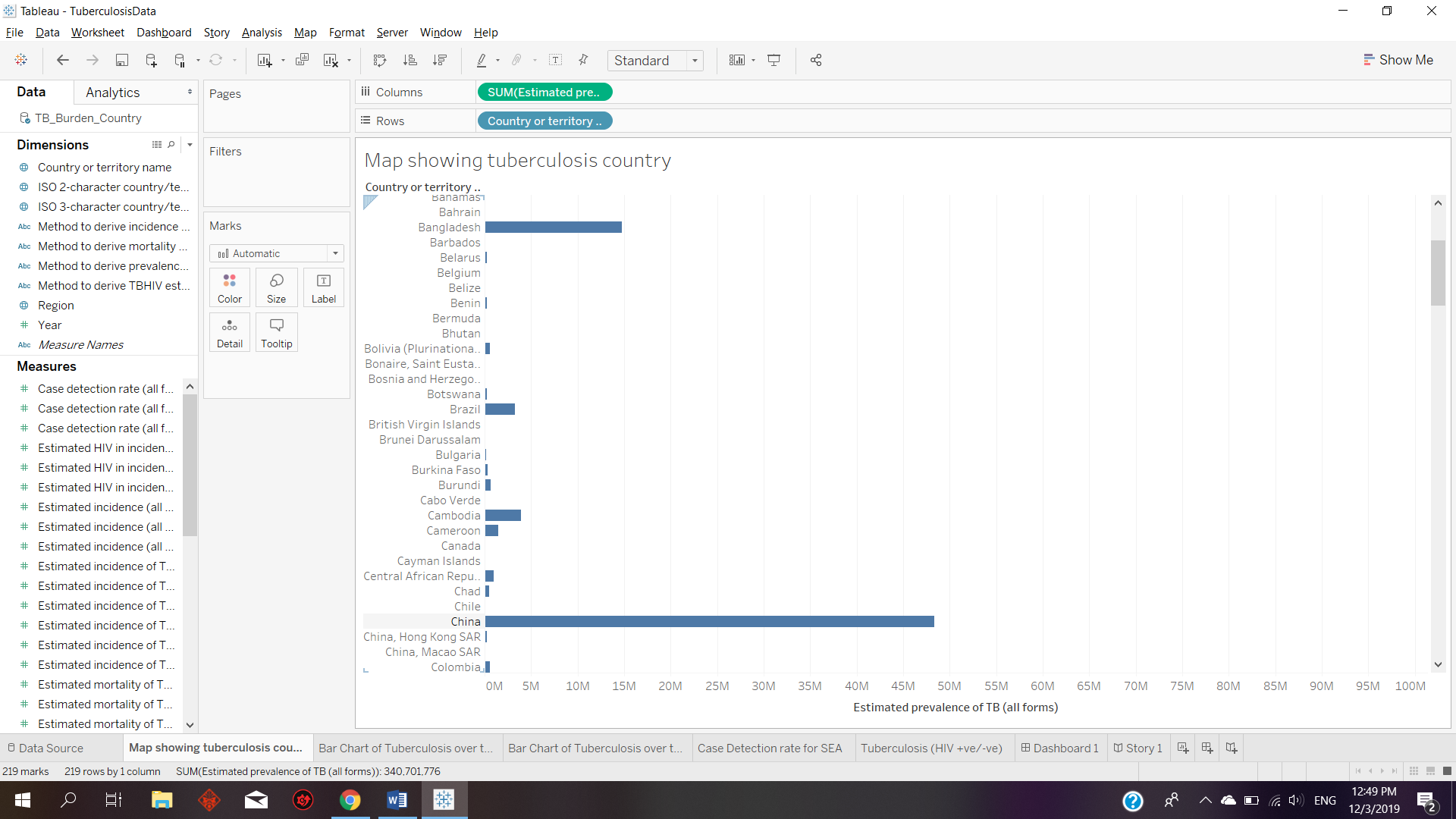


Tableau is also not that user friendly towards beginners as they would have to search online for different tutorials in order to learn. There are too many different functions provided by the software and users would have to perform trial and error in order to try to obtain the intended data visualization design.

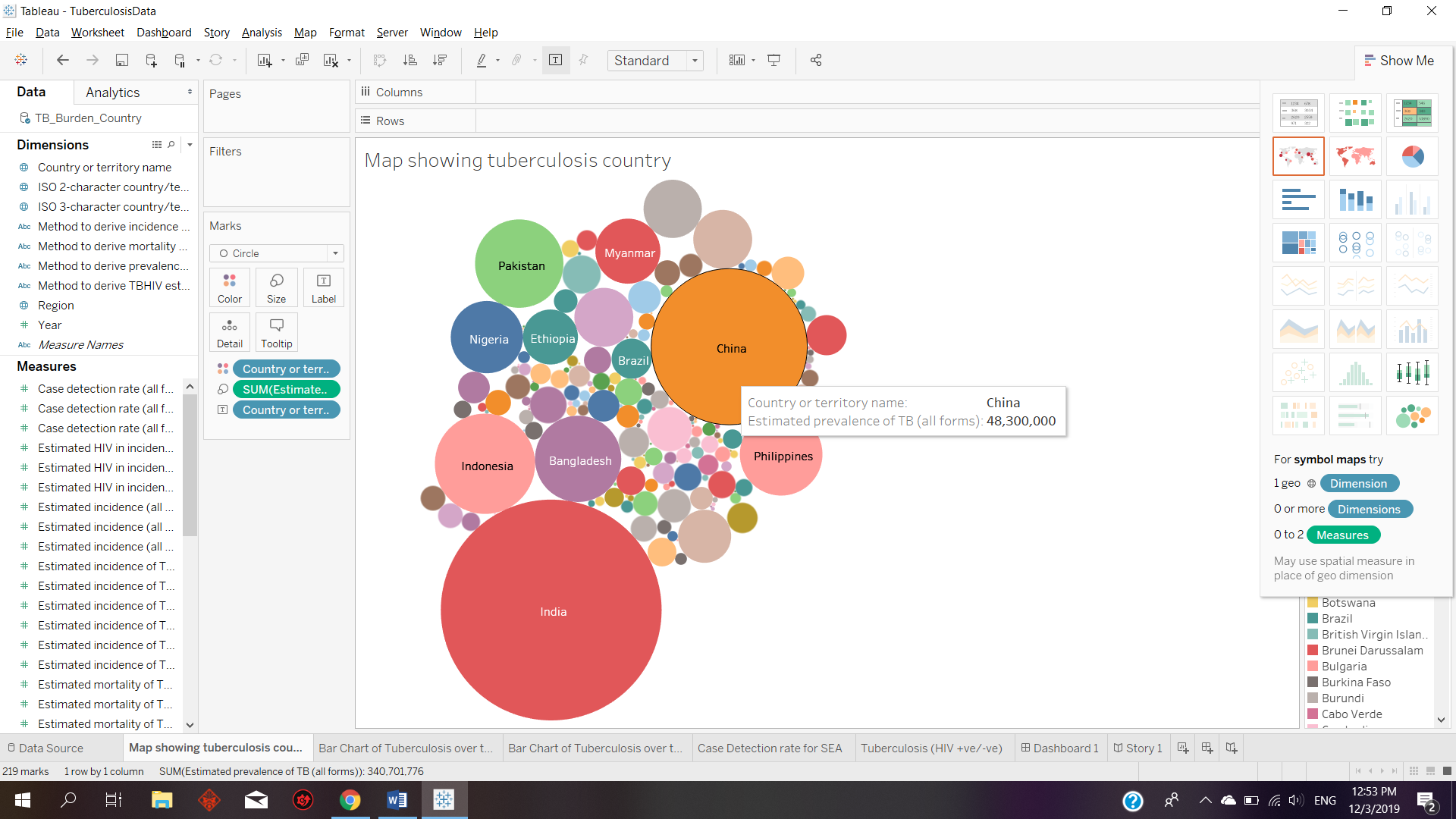
**4.0 Unexpected finding/insight/discovery**

An unexpected finding is that sometimes a vertical bar chart would lose its variables data for example the country China variable name is missing, however when the data set is transposed the value would return.

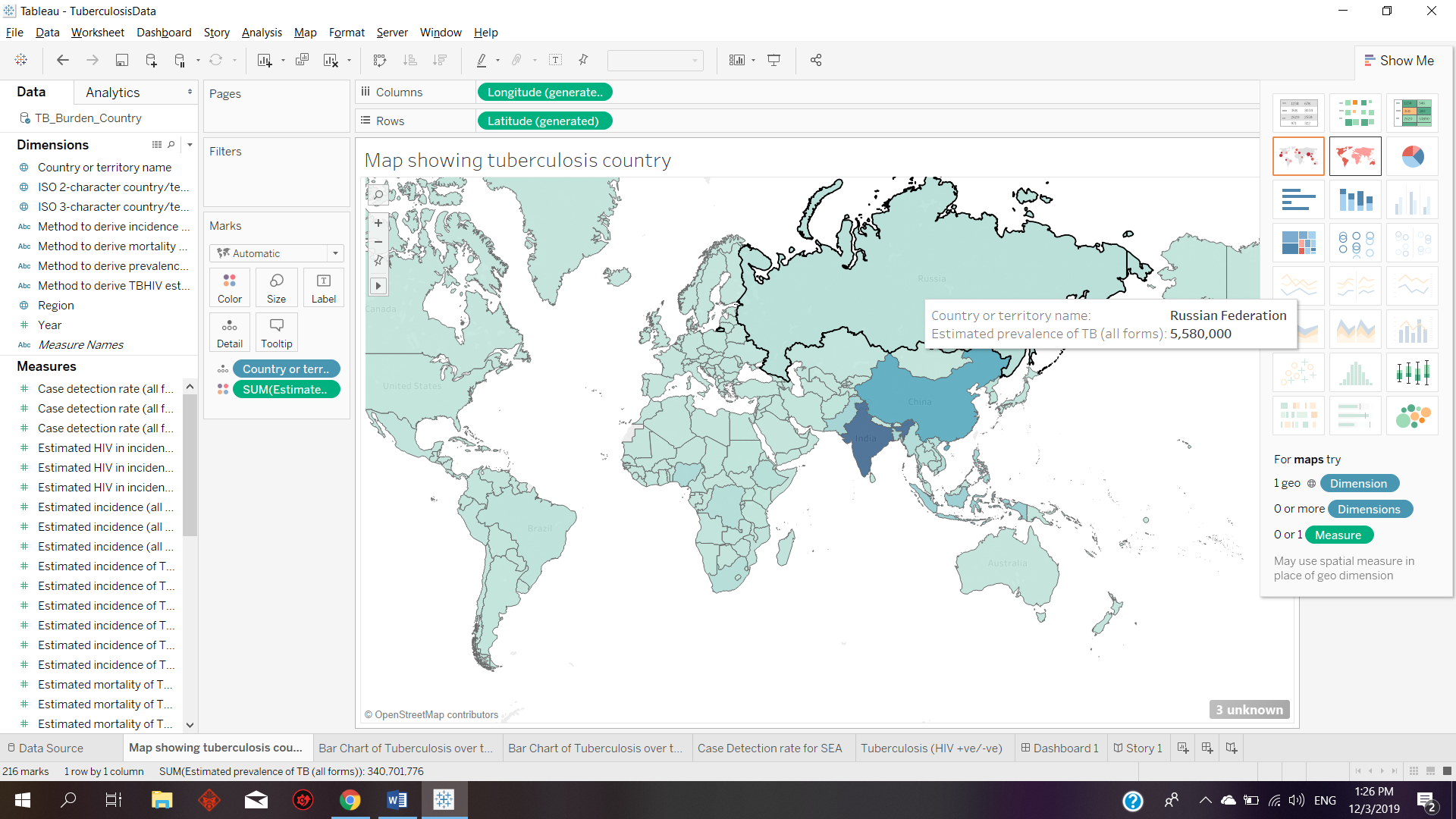




Certain details would also not be shown when there are values where the range are too far apart. For example, most countries name are not displayed in the packed bubble diagram as shown below.



However, my hypothesis is not supported whereby a higher total number of populations would indicate higher prevalence of tuberculosis and countries like Russia has lesser prevalence of tuberculosis compared to other smaller countries such as South Africa.



Based on the system, it is shown that most SEA region countries’ case detection rate are increasing over the years except for India, which was an unexpected finding as India was hypothesized as a more advanced country would have higher case detection rate over the years.



**5.0 Lesson Learnt**

The primary lessons that I have learnt from designing this data visualization would be not to over-plot. Using filter, I can reduce the number of variables to be displayed so that the graph does not look too messy or there would not be missing details. Besides, a good choice of colour is also important to represent the data. Selecting the right visualisation tool is also important as not all graphs are suitable to represent different data. For example, to show data over different country I learnt to use maps to represent as it is fast and easy to view because users can also hover to specific location and each data can be shown. Maps can also contain more data and compact each data to different countries. Not only that, adding labels is also important as it will help to distinct different data from different variables. I have also learnt that tableau can be used to generate dashboards and storyboard in order to help users to present their data.

**6.0 References**

[**https://public.tableau.com/en-us/s/resources**](https://public.tableau.com/en-us/s/resources)