Salome Botelho-homework 1

11/11/17

*A)Three main observations:*

1-From the analyses of the state of the projects taking into account the “Category” in the various countries, we can verify that “theater” is the category with highest number of “successful” and “failed” projects. The category with higher number of “cancelled” projects was “technology”. “journalism” is the category is lowest number of projects and all have been cancelled. Interestingly, US and GB were the strongest contributors with approximately 84% of the total number of projects with a prevalence in the categories of “theater” and “music”.

2-Analysing the sub-categories of the projects, we can verify that “play” is the most prevalent sub-category with the highest number of “successful” and “failed” projects. Interestingly, the sub-categories of “documentaries” and “rock” seem to be quite reliable as they show 100% of successful projects. As expected and already mentioned, US and GB are the main project contributors.

3-By the analysis of the state of the projects throughout the 12 months, we can see more projects failing than succeeding but it seems that successful projects tend to appear towards the end of the year around the month of November. Regarding the goal of the projects, we can verify that the higher the goal, the likelihood to fail and the lower the goal is, the higher the number of successful projects. We can also see that from 2012 there was an increase in the overall number of projects.

*B)What are some of the limitations of this dataset?*

This data set is very comprehensive in the information that gives us, regarding the goal, funding, category & sub-category for all the projects. By analyzing the state of the projects against these variables, we can try to extract the important points or “rules” that are guiding the success or failure of a project.

The simplistic analysis done in the previous exercise, gave us an idea of successful categories and sub-categories, perhaps a good time of the year to create a project and that the dimension of the goals for projects seems to be important. However, we should be cautious in taking these correlations very serious: there might be other factors at play that have not be considered in this data set. For example we do not know anything about intrinsic properties of the teams that are developing these projects and there might be features like the size of the team or experience that may influence more the outcome of the project than the variables above considered.

*C)What are some other possible tables/graphs that we could create?*

We can try to plot the number of projects in function of their state using the filter of duration of the projects. We could also calculate the correlation between for example success of a project with some variable by calculating regression (however I have not been able to do it because many of the variables are non-numerical).