PROGETIIGRI SPONSORSHIP ANALYSIS REPORT

Team members:
Michelle Lukken
Renno Sepp
Kaupo Humal

Our repository:
https://github.com/kaupohumal/IDS_project

1. Business understanding

Estonian educational governmental agency HARNO has provided different schools with IT hardware for 8 years. Thanks to HARNO kindergartens, general education schools and vocational training institutions can apply for support for the purchase of microcontroller development boards, 3D printers, programmable drones and different robotics, electronics and mechanics kits. They have information in Excel format of how many and what products each school has received in that timeframe but the information is not usable in its current state. Our goal is to provide HARNO with usable data that can be easily analyzed and visualize their expenses to different areas and educational institutes. For that, we are going to do data cleaning, processing and find frequent patterns based on what products schools have received and visualize it on a regional level on a map. If possible we can collect additional data on the size of educational institutes and predict what schools would need more funding. Based on the discoveries we make we plan to do some more interesting data visualizations to show our findings in the best way possible. Our results will be measured based on how usable and user-friendly we manage to make data.

Our projects resources will include our team, Python as our main software and HARNOs progetiigri equipment application round information as our data. For this project, we are going to use Python with its libraries Numpy, Scipy, Pandas, Matplotlib, Scikit-learn and possibly others. Since the data is in Excel we are going to do some work there as well and if necessary or a lot easier we would be interested in making an SQL database for our products. Our data consists of different schools with unique id-s, their regions, types, equipment lists and the amount of money they have received separately each year.
The data we will be using is available for everyone at HARNOs website under projects: https://projektid.edu.ee/display/progetiiger/ProgeTiigri+seadmete+taotlusvoor+2021?preview=/81365069/94439710/Seadmete%20taotlusvoor%202014-2021.pdf
Requirements for acceptable finished work include thorough data cleaning with multiple visualizations containing a depiction of regions. Possible risks and setbacks include but are not limited to not finding a lot of patterns in data and oversimplifying data.
There is no difficult terminology in this project, terminology is needed only to understand the IT hardware which can be easily done by googling.

The benefits of this project lie in the reduction of time needed to survey all schools and their products to determine which schools need money for the oncoming applications. Since we are doing this work without getting paid and we do not spend anything else but our time on this project, the benefits outweigh the expenses.

With our data mining we wish to accomplish a deeply cleaned data that is easy to read and visualize. Addition to that we wish to provide the business some practical knowledge on which they can rely on making their future decisions. All in all the set goal is to accomplish the tasks set by the task provider (Extract from (Seadmete nimekiri) column how many of each type of products schools have received, visualize money spent on maps of Estonia for yearly and total spent at county (region) level).

2. Data understanding

The first step of Data understanding is gathering data. For many projects we believe this is an essential task, as it is the base on which you answer stated research questions, test hypotheses, and evaluate outcomes. For us this part was mostly done, as the task provider also gave us access to the data he had collected over 8 years about the support given to institutions during different periods. As our goal is to create visualizations containing a depiction of regions and other interesting finds from the given data, we have a clear path set by the task provider. We thought about finding more data from the internet, but as the request for the outcome was very concrete and the given data actually demanded all our needs we did not find it necessary to add some data just for the reason of adding it.

The next step is to describe the data to give a better understanding of what we have and will it be the information we need. As said before this data is provided to us from HARNOs website. To get a better format of the information we were able to get an Excel of the same data. The dataset consists of 692 different kindergartens, general education schools and vocational training institutions as the cases that have gotten any support from HARNO in the last 8 years. The total amount of fields is 13 and they describe the following: 1) ID of the institution, also the key, 2) The name of the supported institution, 3) The county of the institution, 4) The type of the institution (kindergarten, basic school, kindergarten/basic school, gymnasium or adult school), 5) The list of devices provided. (The field is full of unclean data, written in different logic. All the devices are followed by the amount all in one field.) 6-13) Represent years from 2014-2021 each having a value of the support in euros. So each year they add a column with the next year's number and start adding the support amounts, leaving the field empty if the institution did not recieve anything that year. This data provided needs to be cleaned, but after that provides perfectly the information we need to accomplish our main goal of creating a Estonian map with the money spent on different counties in different years.

We thoroughly covered some steps of exploring data in the previous (Data description) block but let's briefly look them over again. We already talked about the fields and the possible values of the fields. One of them needs a bit more attention, the field of devices and the amounts 5). As said before all the devices are put together into one field, with no particular logic. Also if an institution has gotten support multiple times, all the devices are still put together in one field. There seems to be some pattern of separating the devices supported in different years by ENTER and most of the products are separated with comma or semicolon. There still stands a problem that some of the product names also use commas, so if we need to find any specific information about the devices, there will be a great task of fully cleaning the field.

For the last step we need to verify data quality. We have a data set with over 600 different institutions ever gotten help from the support program. Most of the data is well organized and essential for us to reach our goals. Nevertheless there still is a field, which is stacked with information and needs to be dealt with in order to get useful information out of it. Certainly it takes time, but seems to be managable. It is possible that we lose some information in the process, but as our main goals do not exactly need the results of the field, it is not essential that all the information stays. Alternatively we can take contact with the dataset provider to clarify some details, as we already have done before.

3. Plan of our Project

Task nr.	Description	Michelle	Kaupo	Renno
1	Clean the dataset (Seems to be a big task that might take a lot longer than estimated here)	6h	6h	6h
2	Extract from (Seadmete nimekiri) column how many of each type of products schools have received (1st of our 2 main tasks)	5h	5h	5h
3	Visualize money spent on maps of Estonia for yearly and total spent at county (region) level (2nd of our 2 main tasks)	5h	5h	5h
4	Other cool data visualizations (We try to find more interesting patterns that might help the firm make better decisions.)	5h	5h	5h
5	Visualizing the results (As most of the results we will get need to put on a map, it might take time to get the best visualisation out of it)	3h	3h	3h

6	Preparations and presenting of the project	6h	6h	6h

We will try to work on everything together in order to get a better understanding of the projects every step.

We will be mostly using Jupyter notebook and Python with its libraries in the reasoning of using the skills we have learned during the course.

We are not fully certain about the methods we will be using, but we will note them in the future with an accurate description of how we did it.