D8 report

Task 1

https://github.com/Katillus/IDC_Project_Fires_in_Estonia

Task 2

Identifying your business goals

Background

Our main goal with this project was to find a topic that is relevant in real life. We browsed through Estonia's open government data portal and found a dataset about building fires. We chose the project, because we think it is an important topic and we couldn't find any reports about building fires, that analysed more than a year's worth of data at a time. Our dataset hasn't been analysed before and we think that the results could be useful in real life. There's more than 1000 building fires in a year and the reasons behind the fires vary a lot.

Business goals

Our main goal is to find out which counties have the most fires in a month and why. We also want to see if we can find any patterns. For example if a county has a lot of fires that have happened because of the heating system or the incorrect installation of a device or system, then that could tell us the county has a lot of unsafe buildings. That could be very dangerous and this info could be relevant for people living in similar buildings in the county. It would be helpful for the county to be informed about those dangers so it would be possible to take it into account while planning a budget to either fix buildings or decide whether it's sensible to tear down some buildings.

Business success criteria

We will know if the project has been a success when we can accomplish our goals. We have two main goals and one additional and if we accomplish the main goals, then we know we have already been successful. If we can get results with the third goal as well, then that would be really good, but it isn't necessary for the project's final results.

Assessing your situation

Inventory of resources

We have three people working in the team: Karolin, Kati and Geitrud. If needed, we can ask for help from our supervisors as well. Our project is based on one dataset, which consists of all the building fires reported in Estonia over the last 7 years and it shows all the fires by county, year, cause and month. The data is in a .csv file and we will use Jupyter Notebook for the data-mining.

Requirements, assumptions, and constraints

Results of projects will be presented in the poster session on Thursday, December 17, 2020 at 14:00-17:00. We have to be finished with the project by then and also make a poster about the project. The requirements are that we have to demonstrate mastering some topics from the Introduction to Data Science course and in the readme-files of our code repository page we must specify the origin of the data and provide a short description of the data. The data doesn't have to be public as long as the instructors have access to it.

Risks and contingencies

A risk with every group project is that the workload could be distributed unevenly as we can't be certain that all the tasks take the same amount of time. Slacking is another risk in a group project, but we are certain it won't be the case in ours. Also because of corona, we most likely can't meet up, which makes the completion of the project more difficult.

Terminology

- Data Visualization - The art of communicating meaningful data visually. This can involve infographics, traditional plots, or even full data dashboards. (https://www.dataquest.io/blog/data-science-glossary/)
- Clustering - Clustering techniques attempt to collect and categorize sets of points into groups that are “sufficiently similar,” or “close” to one another. (https://www.dataquest.io/blog/data-science-glossary/)
- Variance - The variance of a set of values measures how spread out those values are. (https://www.dataquest.io/blog/data-science-glossary/)
- K-means clustering - an unsupervised clustering algorithm that gathers and groups data into k number of clusters. Clusters the unlabelled points into groups by analysis of the mean distance of points. Needs unlabelled data to train. (https://pythonprogramminglanguage.com/how-is-the-k-nearest-neighbor-algorithm-different-from-k-means-clustering/)
- KNN - represents a supervised classification algorithm that will give new data points accordingly to the k number or the closest data points. Needs labelled data to train on. (https://pythonprogramminglanguage.com/how-is-the-k-nearest-neighbor-algorithm-different-from-k-means-clustering/)
- Random Forest algorithm - a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. (https://builtin.com/data-science/random-forest-algorithm)
- Decision Tree classifier - It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees. (https://en.wikipedia.org/wiki/Decision_tree_learning)
- Ridge - simplifies models, but instead of eliminating features entirely it minimizes their effect. (https://towardsdatascience.com/regression-and-its-variants-simple-multiple-lassso-ridge-and-stepwise-regression-9c144fd7a7b2)
- Lasso - simplifies regression equations that have too many features by completely eliminating some of them.

(https://towardsdatascience.com/regression-and-its-variants-simple-multiple-lassso-ridge-and-stepwise-regression-9c144fd7a7b2)

Costs and benefits

The project will not cost us anything, but the Estonian Rescue Board could benefit from our project. When we have finished with the project, then we will see if our results are useful to them and if they are, we will send the project to them.

Defining your data-mining goals

Data-mining goals

We have two main goals and one additional. First, we will analyse how the amount of building fires in Estonia varies by county and month, which could tell us which counties have the most fires and why. We will also want to see if there are some kind of patterns, for example when we also take months into account, we could see if seasons have an impact on the number of fires. Our second goal is to make a cluster model of fires by similar causes so the prevention work could be more effective. We looked at the different causes and some of them are quite similar: for example there's a cause about open fire and also a cause about children using open fire. We had one more goal, but we are not sure if we can get actual results with it, but we will try. The goal is to create a prediction model that would predict next year's fires by county, month and cause. The results could actually be useful by redistributing fire brigades by county so there would be extra brigades ready in some counties in the most critical months.

Data-mining success criteria

We have a successful project when we have analysed the data and made cluster models. Estonias Rescue Board has made diagrams about the data (but only one year at a time), so we can later check if our results are similar.

Task 3

Gathering data

Outline data requirements

The required data will be in .csv format and that is available for download right from the resque.ee webpage. In the .csv format it will be possible for us to start working with the data without converting it into some different type. The time range will be from 2014 to 2020 since this is the longest available time range for acquiring data about this topic from this given webpage.

Verify data availability

The required data does exist on the webpage we found it from. We have already tried to download it and we have verified the data is usable and publicly available.

Define selection criteria

The data will be solely downloaded from the webpage https://www.rescue.ee/et/hoonetulekahjud. All the available months, years, counties, and determined reasons of fire which will are brought out on the downloaded .csv file format, will be relevant to this project.

That means months from January to December (with the exceptions of 2020 December since it is in the future and January and February of 2014 since data from those months isn't included into the downloadable file for unknown reasons), all the years from 2014 to 2020, all Estonian counties, all possible reasons and all quantities of fires by county and month will be taken into account.

Describing data

There are 6 columns in the data with headings: "Maakond" a k a County, "Aasta" a k a Year, "Hoone liik" a k a Type of building (either residential or non-residential building), "Tekkepõhjus" a k a Cause, "Kuu" a k a Month, "Hoonetulekahjud" (number of building fires). There are some null values in the data. Those are marked as "!Sisestamata!" and it will be our job to change those into something more suitable for marking non-existant value.

There are 5691 rows of data in total (without taking the column names into account). The data is distributed first by year (the 2014 data is at the beginning of the file and 2020 is at the end), then by month (January marked as 1, December as 12).

Exploring data

Possible values for “Maakond” a k a County: 'Valga maakond', 'Tartu maakond', 'Rapla maakond', 'Lääne maakond', 'Harju maakond', 'Võru maakond', 'Viljandi maakond', 'Saare maakond', 'Pärnu maakond', 'Põlva maakond', 'Lääne-Viru maakond', 'Järva maakond', 'Jõgeva maakond', 'Ida-Viru maakond', 'Hiiu maakond', '!Sisestamata!'.

Possible values for “Aasta” a k a Year: 2014, 2015, 2016, 2017, 2018, 2019, 2020.

Possible values for “Hoone liik” a k a Type of building: 'Mitteeluhooned', 'Eluhooned', '!Sisestamata!'.

Possible values for “Tekkepõhjus” a k a Cause: '!Sisestamata!', 'Laste mängimine lahtise tulega', 'Rike elektripaigaldises', 'Suitsetamisel', 'Kuritahtlik', 'Summutist jt seadmetest lenduvad sädemed', 'Rike elektriseadmes', 'Lahtise tule kasutamisel', 'Kindlaks tegemata põhjus', 'Rike kütteseadmes', 'Konstruktsioonipuudus', 'Muu hooletus', 'Kütteseadmete kasutamisel', 'Tahma süttimine suitsulõõris - tahmapõleng korstnas', 'Kulu põletamine', 'Elektriseadmete kasutamisel', 'Tehnoloogilise protsessi teostamisel', 'Toiduvalmistamisel (kõrbemine)', 'Tahma süttimine suitsulõõris - tahmapõleng, tekkis kahju', 'Mootorsõiduki elektri- ja toitesüsteemi rike', 'Tuletöödel', 'Teadmatus', 'Tehnilise seadme rike', 'Pikselöök, keravälk', 'Isesüttivate ainete ja materjalide hoidmisel', 'Muu ebaõige käitumine', 'Seadme või süsteemi vale paigaldus','Loodusnähtused'

Possible values for “Kuu” a k a Month: 1-12

Possible values for “Hoonetulekahjud” (number of building fires): 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 22, 23, 24, 27.

Verifying data quality

There are NaN values in columns “Maakond”, “Hoone liik” and “Tekkepõhjus”.

The NaN value occurs in the column “Maakond” 5 times. Out of the total 5691 values it isn’t really significant proportion and will probably not affect the outcome.

The NaN value occurs in the column “Hoone liik ” 11 times. Out of the total 5691 values it isn’t really significant proportion and will probably not affect the outcome.

The NaN value occurs in the column “Tekkepõhjus” 131 times. This is about 2.3% of values from the total amount of values for this column. This is a little bit alarming amount but there’s nothing we can do to reduce this and all the other causes of fire give us plenty of information to still make helpful summaries of this topic so it isn’t really enough of a reason to reject this dataset. Furthermore, we are concentrating on finding counties with the highest amount of fires and we have already determined it’s possible to bring out which are the more common causes for fires.

Task 4

Detailed plan

Our tasks can be divided into 5 categories:

1) Preparation and introduction (pink)
2) Analysing the data (purple)
3) Clustering (blue)
4) Creating a prediction model (mint green)
5) Summarizing and presenting (green)

List of tasks and workload in hours for each team member can be seen in the table below.

Task	Geitrud	Karolin	Kati	Total
Finding a topic and setting goals	4	2	2	8
Preparation for the introduction	1	1	1	3
Presenting in the practice session	1	1	-	2
Analysing the data	4	4	4	12
Data to dataframe in Jupyter notebook	1	-	-	1
Visualizing and plotting the data	-	2	2	4
Clustering (using different methods)	5	5	5	15
Creating a prediction model	7	7	7	21
Summarizing	2	2	2	6
Making the poster	4	4	4	12
Making the video	1	3	3	7
TOTAL	30/30	31/30	30/30	91/90

Methods and tools

We are using different methods we have learned from the course to analyze, visualize, cluster and make prediction models.

For visualisation we will use different plot-types we have learned in this course. For clustering we will use the K means algorithm. This can be quite time-consuming since we plan on doing it on a map of Estonia. We will use different classification and regression models, for example Random Forest and Decision Tree classifiers, as well as multi-variate linear regression models (simple linear, ridge and lasso). During the project we will decide which of those we are going to use, because we are not sure if we can predict anything from our data yet (as we have stated before).

We will do our project on Jupyter Notebook.