Task 1. Setting up (0.5 points)

- Add reports from following subtasks (business understanding, data understanding and planning) to the repository as a single separate PDF file named GROUP_NR_report.pdf (e.g "A0_report.pdf").
- Add the link of the repository also to the report.

Group C11
Link to the GitHub repository: **https://github.com/mannapuder/IDS2020-kaggle-uswildfires**

Task 2. Business understanding (1 point)

NB! Don't forget to mention your project title and team members at the beginning of the report.

Developing a business understanding within CRISP-DM consists of four tasks: identifying your business goals, assessing your situation, defining your data-analysis, data-mining or machine learning goals and producing your project plan. For this exercise, please, develop a business understanding of your project. According to CRISP-DM, you should report the following:

- Identifying your business goals
    - Background
    - Business goals
    - Business success criteria
- Assessing your situation
    - Inventory of resources
    - Requirements, assumptions, and constraints
    - Risks and contingencies
    - Terminology
    - Costs and benefits
- Defining your data-mining goals
    - Data-mining goals
    - Data-mining success criteria

Please, follow this given structure and cover all these aspects in your report. Consult this PDF-file with a chapter on Embracing the Data-Mining Process for more information on each of the deliverables. Keep the report concise and feel free to state that some aspect is not relevant in your project. If your project is not meant to benefit a 'business', then please specify who will benefit from the project and perform business understanding from their perspective. For instance, this could be either one or multiple

individuals, organizations, or societies. Please focus on the goals that you plan to directly contribute to, not on the generic goals (like making the world a better place).

The report of task 2 should be 400-800 words.

Identifying business goals

Background

A fire is considered a wildfire if it is unexpected and occurs in a natural area such as forest or grassland. Wildfires can have multiple natural causes, for example lightning, but it is increasingly common that these are caused by human activity. In most cases it is actually not known how exactly the fire started.

Wildfires can have a huge impact on local citizens' lives by causing serious air pollution, vegetation (crops) and property destruction. It is estimated that worldwide approximately 339,000 people die due to the effects of wildfire smoke each year and because of climate change the size and frequency of the fires is increasing.

The goals of this project

This project is focusing on the wildfires in the United States that occurred from 1992 to 2015. Our main goals are as follows. Firstly, by analyzing the data about the wildfires that occurred in the past, we can estimate whether these have become more or less frequent over time and use the results to form predictions about the future as well. This can be useful for the locals to evaluate fire danger in the future and possibly to the other researchers as well. For example, the results of this analysis can be combined with similar works about other areas in the world and therefore be used to create predictions about global warming.

Secondly, by analyzing the locations of known fires, we will create a map which estimates which areas are the most and which are the least fire-prone. This can be useful for local authorities and citizens in order to evaluate the fire danger in their area. It has the potential to make people more aware of the possible danger which hopefully leads to more careful behavior that can reduce the number of fires caused by human activity.

Thirdly, we will create a machine learning based model which predicts the possible causes of the fires. As most causes remain unknown, the estimations can still be used in fire prevention in order to minimize the risk factors, especially when these are caused by human activity. Two additional maps will be added showing the fire causes specific to different areas. One map for natural causes and the other map for fires caused by human activity.

Success criteria

The success of the project can be measured by analyzing the data about the years greater than 2015 and possibly the data that might be collected in the future. The project can be considered

successful if the predictions about the future turn out to be accurate enough to make decisions in fire prevention based on them.

Assessing the situation

Inventory of resources

- "1.88 Million US Wildfires" dataset from Kaggle which is about wildfires in the US that occured from 1992 to 2015.
  Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA*FOD*20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4
- USHCN daily weather data from the years 1992-2015
  Menne, M.J., I. Durre, B. Korzeniewski, S. McNeal, K. Thomas, X. Yin, S. Anthony, R. Ray, R.S. Vose, B.E.Gleason, and T.G. Houston, 2012: Global Historical Climatology Network - Daily (GHCN-Daily), Version 3.26
- 3 team members

Requirements, assumptions, and constraints

The deadline of the project is Thursday, December 17, 2020. The assumed amount of work is 90 hours in total (30 hours per team member).
Both datasets can be used without additional permissions or fees.

Risks and contingencies

This is a very short project and therefore it does not involve any serious risks.

Terminology

The topic is simple for a regular person to understand. We haven't encountered any specific terminology so far. It will be added on the fly if necessary.

Cost-benefit analysis

Isn't relevant for this project. This is a very small and short project that involves zero costs.

Data-mining goals

- Reduced datasets containing only the data relevant to our project (see the selection criteria in data understanding report)
- A map about the most and least fire-prone areas in the US.
- Two additional maps showing the causes of the fires by area (one map about natural causes and the other about human activity)

- A model based on machine learning which predicts the possible causes of the fires that have already happened.
- A model which aims to predict the frequency of wildfires in the future.
- A poster presenting the findings of this project with visualizations of the data and predictions. The aim of the visualizations is to give a good overview of the changes occured over the years.

Data-mining success criteria

Our aim is to predict the frequency of future wildfires with at least 90% accuracy and the causes of the already occured fires with at least 70% accuracy. The maps should cover the whole country.

Task 3. Data understanding (2 points)

Data understanding within CRISP-DM consists of performing four tasks: gathering data, describing data, exploring data and verifying data quality. For this exercise please develop a data understanding of your project. Report the results of the tasks according to the following structure:

- Gathering data
    - Outline data requirements
    - Verify data availability
    - Define selection criteria
- Describing data
- Exploring data
- Verifying data quality

Consult the above-given book chapter to understand what is expected under all these deliverables. Take inspiration from when describing and exploring the data. As a result of this exercise, you should have gathered and understood the data. You should have decided which parts of the data you are potentially going to use and understood the meaning of all fields within these parts. Note that data cleaning is part of the data preparation step in CRISP-DM but you might choose to do some of it already during this task.

The report of task 3 should be 400-800 words.

Gathering data

Data requirements

For our data mining goals it is important to have accurate data about wildfires in the US in recent history (time, location) and their causes. We would also need some additional data such as weather data from US weather stations close to the fires from the times of the fires.

Data availability

The required data is publicly available.

https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/ (ghcnd-stations.txt and gchnd_hcn.tar.gz)
https://www.kaggle.com/rtatman/188-million-us-wildfires

Selection criteria

We will use the “1.88 Million US Wildfires” dataset from kaggle, more specifically the “Fires” table from the SQLite database file given there. The fields in that table that are of interest to us are “FIRE_YEAR”, “DISCOVERY_DOY”, “STAT_CAUSE_CODE”, “STAT_CAUSE_DESCR”, “LATITUDE”, “LONGITUDE” and “STATE” as well as “DISCOVERY_DATE”
Additionally, we plan to use the USHCN daily weather data from the years 1992-2015 to…
There, we would be mainly interested in “TMAX”, “TMIN”, “TAVG”, “PRCP”, “WT03” and “WV03” and at least the quality flags for each. (“Weather Type” and “Weather in the Vicinity”, 03 means lightning). We’d only need the data from the days a wildfire happened from the station(s) closest to the wildfire.
For connecting the two datasets, we also need the coordinates of each of the stations, which are given in a text file with the weather data.

Describing data

Source: given above.
Format: SQLite database file for the wildfires, fixed-width files for the weather data
Number of cases: 1880465 wildfires, weather data of 1215 stations
Descriptions of fields:

- The wildfires data:
    - FIRE_YEAR: calendar year of the fire as an integer
    - DISCOVERY_DOY: the day of the year that the fire was discovered as an integer
    - DISCOVERY_DATE: (Julian) Date on which the fire was discovered or confirmed to exist.
    - STAT_CAUSE_CODE: the numerical code for the statistical cause of fire
    - STAT_CAUSE_DESCR: the description of the statistical cause of fire
    - LATITUDE: Latitude (NAD83) for the point location of the fire.
    - LONGITUDE: Longitude (NAD83) for the point location of the fire.
    - STATE: Two-letter alphabetic code for the state in which the fire burned
- The weather data:
    - Station ID: the ID of the weather station
    - Year: the year of the weather data
    - Month: the month of the weather data

    - Element: the weather element being measured:
        - PRCP: Precipitation (tenths of mm)
        - TMAX: Maximum temperature (tenths of degrees C)
        - TMIN: Minimum temperature (tenths of degrees C)
        - TAVG: Average temperature (tenths of degrees C)
        - WT03: Weather type “lightning” (1 if true)
        - WV03: Weather in the vicinity “lightning” (1 if true)
    - Each element value column represents a day
- The information about the stations:
    - Station ID. the ID of the weather station
    - LATITUDE: the latitude of the station (in decimal degrees).
    - LONGITUDE: the longitude of the station (in decimal degrees).

Exploring data

Wildfire data:

- FIRE_YEAR: no values outside of the expected range and no missing values. The average number of fires per year is approximately 78350 with a standard deviation of roughly 12760. The highest number of wildfires was recorded in the year 2006.
- STATE: There are 52 “states” represented in the data, Washington DC and Puerto Rico being the areas included in this data that are not counted among the 50 official states. No missing values in the data. There number of fires understandably varies greatly between states, with the lowest number of fires in the data being 66 in Washington DC and the highest number of fires being 189550 in California.
- STAT_CAUSE_DESCR and STAT_CAUSE_CODE: No occurrences of description and code not matching. Most of the fires were caused by human activity, mainly debris burning.
- The cause of the fire is missing or undefined in 166723 (roughly 9%) of the cases.
- DISCOVERY_DOY: no missing values and no unexpected values. Roughly half of the wildfires happen in the first and roughly half in the second half of the year (the 50% quartile is on day 164).
- LATITUDE and LONGITUDE: no missing values, no obviously wrong values.

About weather data:

- PRCP: 202862 missing values of the roughly 9.5 million rows, 5239 of the existing values are noted as having failed a quality check. 6.75 million rows of data, so well over half, show no precipitation, while there are some rare occurrences of over 30cm precipitation in a day, however none of the values reach American records, so none are obviously wrong.
- TMAX: 244818 missing values and 90140 values flagged for failing a quality check. Of the 99 cases where the TMAX was below -30 degrees, 48 failed a quality check and only one recorded maximum temperature was below -40 degrees, which failed a quality check. 202 values were between 50 and 54 degrees, which is incredibly close to

America's record, but not beyond it, while one value of 56.1 degrees failed a quality check and one value of 5537.2 is obviously outside the realm of possibility.
- TMIN: 253178 missing values and 127348 quality check failures. All values below -47.8 degrees failed a quality check, as did all values above 41.7 degrees. A minimum temperature above 35 degrees does seem somewhat extreme, but might still be possible.
- TMIN and TMAX: there are 4227 cases where TMIN is higher than TMAX, though in all those cases at least one of those has failed a quality check. In 1947 cases TMIN and TMAX are equal and in 342 of those neither TMIN nor TMAX have failed a quality check.
- TAVG: approximately 9.32 million missing values and 254 of the 200147 existing values failed a quality check. Lowest average temperature is -32.2 degrees with no failed quality flag, while the highest is 150 degrees, which is obviously an error.
- WV03: 4495 existing values with no quality flags.
- WT03: 233405 existing values with no quality flags.
- WV03 and WT03: only 30 cases exist where WV03 has a value while WT03 does not

Verifying data quality

Most of the data is well represented. Missing values for the causes of the fires don't really matter when simply looking at the frequency of fires over time, though can be a bit detrimental when trying to predict the causes - could be removed from the data or actually used to see what the final model would predict for them? Most of the weather data has only a few missing values and quality check failures, which can maybe be simply discarded, need further work with the data to be sure. TAVG has a lot of missing values, thus it would be difficult to use it for training a model, so maybe we will need to leave it out of our data, or maybe we can feasibly replace the missing values with the mean of TMAX and TMIN for that day. WT03 is also present in only a small amount of rows, but could be tested to see if including it would improve the model's prediction. WV03 however features in less than 1% of the weather data and thus is probably not very useful and can be simply left out.

Task 4. Planning your project (0.5 points)

Please perform the following tasks:

- Make a detailed plan of your project with a list of tasks. There should be at least 5 tasks. Specify how many hours each team member is going to contribute to each task.
- List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

The report of task 4 should be 100-300 words.

Project plan

Task description	*Assignee*	*Result*	*Estimated time*
Planning the project	All (5h each)	The current report	15h
Selecting and cleaning the data. Detecting outliers and removing them	Kaarel	Two datasets (wildfires and weather) with only necessary columns and quality data	10h
Merging the data	Kaarel	Two datasets joined together	6h
Visualizing the data in order to get the overview of it	Anna	Plots to be included in the final report/poster	6h
Forming a prediction about the frequency of the fires in the future	Anna	Plots with predictions for the upcoming years (similar to the previous task)	7h
Creating the map of fire-prone areas	Anna	A map which shows which areas are the most and which are the least fire-prone, included in the final report/poster	4h
Creating 2 additional maps of the wildfires.	Anna	Two additional maps of the wildfires which indicate the causes. One is about natural causes and the other is about the causes involving human activity.	4h
Data preparation for modelling	Andre	Second set of the same data suitable for use in machine learning models	6h
Model selection and testing	Andre	A machine learning model which works the best when predicting possible causes of the fires	7h
Fine tuning the model	Andre	The best set of parameters for the previously chosen model that increases the accuracy the most	5h
Evaluating and correcting the results	Andre (4h), Kaarel (5h)	Verification of the tasks done so far and some conclusions whether the goals have been achieved or not. Corrections if necessary	9h
Writing the final report	All (4h each)	The final report/poster about the	12h

		project and results	
Presenting the project	All	Presentation	0.25h

Methods and tools we plan to use

- Python programming language
- Pandas for data manipulation and analysis
- Spyder IDE for building the models and predictions
- Jupyter notebook to present the final results
- Matplotlib for generating the plots
- Seaborn for the heat maps
- Machine learning methods from scikit-learn library - the exact methods are yet to be determined. We will try which works the best.