Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática

Business Intelligence

Forest Fire Prevention and Management

Teachers

Catarina Helena Branco Simões da Silva

Bernardete Martins Ribeiro

Team

Filipe Miguel Fonseca dos Santos, nº2017271196

Ricardo David da Silva Briceño, nº2020173503

Index

Part 1 – Context

Nowadays, our forests have been extremely harmed by fires, which have in consequence caused the destruction of millions of hectares of the ecosystem as well as many agricultural lands. The fires have various causes, which makes their analysis very complex. However, human intervention is one of the main causes of starting and propagating fires. As such, this is a topic that demands a lot of attention.

2.a Companies and Institutions

In this case the companies involved will be the "Serviço Nacional de Bombeiros (SNB)", "Instituto da Conservação da Natureza e das Florestas (ICNF)", "Serviço Nacional de Proteção Civil (SNPC)", the Government itself, the "Direcção Geral de Recursos Florestais (DGRF)" and the "Guarda Nacional Republicana (GNR)".

The SNB is included since Local Firemen are the main line of defence against these threats. The ICNF has multiple plans and protocols related to protecting the forest from fires. The SNPC acts in accordance with a plan of the previous institution. That plan is the "Plano Nacional da Defesa da Floresta Contra Incêndios (PNDFCI)", which was approved by the Government. This plan lead the government to understand the importance of taking care of this objective in a simple and objective manner. The DGRF is in control of the structural prevention of fires. Meanwhile the GNR takes care of patrolling and detecting new fires.

2.b Business

Our business would be a service for fire prevention and management. It would allow the DGRF and the GNR to get better predictions of where and when fires are more likely to happen, as well as how dangerous they can potentially be. This business would be based on Portugal.

2.c Reasons for the Success of the Companies and Institutions

These companies and institutions are essential in keeping people safe from the various dangers that fires can cause. As such, they are considered paramount for the existence of a society.

2.d Main challenges the Companies and Institutions face

The main challenge present is the need to very efficiently manage all the resources, since they are various. There needs to be a good management of the water used, the teams sent for fighting the fires, the vehicles used by the firemen (which can also be aerial vehicles) and the identification of safe locations for the firemen to act, as to reduce the risks.

3. Data source

3.a Location of the Data

https://www.kaggle.com/sumitm004/forest-fire-area

3.b How the Data will be accessed

The source mentioned has a downloadable .csv file, which we converted to the .xlsx format. With this format we can easily access all the rows and columns, which are clearly identified, which makes working with them a more natural process.

3.c Amount and Quality of the Data

The data in the source refers to the time period of January 2000 to December 2003 and has 517 records. It was collected in the Montesinho Natural Park in "Trás-os-Montes", which is located in the northeast region of Portugal. The xlsx file in question has a size of 46KB. In this case, the data of the source will not be updated, so there won't be a need to update the analysis.

3.d Potential issues with the quality of the data

A limitation is the fact that the data refers specifically to the Montesinho Natural Park. As a result, it might not be fully representative of the situation of the entire country of Portugal.

Another potential issue is how old the data is. This could lead to the data not being as meaningful as it could be if it were more recent.

The total size of the dataset is also small (33KB), which can make the analysis end up not being as representative as expected. On the other hand we are considering the subject of fires, specifically in Portugal, so it would be very difficult to find a bigger dataset with the same level of quality.

It's worth noting that the dataset only mentions the day of the week and the month, but not the year. This could hinder the potential of the analysis.

4 Goals of our BI Solution

4.a/b/c How the BI Solution can help the Companies and Institutions thrive

Our BI solution can give the mentioned Institutions many advantages. For example, it can help predict whether a fire had natural causes or if it was started intentionally by humans. We can try to do this by analysing the temperature at the time of the fire. This would be of great interest to the GNR, who need to know if they need to intervene.

We can also analyse which areas are more likely to be affected by fires (which ones have had a bigger burned area), so that these can be given more attention by the corresponding Institutions (GNR and the GDRF).

It should also be possible to predict when fires are more likely to happen and how dangerous they could become (with an analysis of the FFMC index), which can lead to better management of the resources necessary for taking care of them. This could be useful for both the DGRF in the prevention of fires and the SNB so they can know what kind of issue they are facing.

Regarding the mentioned institutions, the people inside them who would be interested in seeing the results of the analysis would be the management responsible for taking care of these issues, as well as making decisions based on that analysis.

4.d Most Relevant Questions we can find answer for

1. "In which places and moments are fires more likely to occur?"
2. "How much impact do the months have on the number of fires?"
3. "Does the day of the week have a significant influence on the occurrence of fires?" (This would suggest intentional fires)
4. "How much do meteorological conditions affect the spread of fires?"
5. "How accurate are the Canadian Indexes used in fire prevention?" (FFMC - Fine Fuel Moisture Code, DMC – Duff Moisture Code, DC – Drought Code)
6. "How much does the wind affect the Initial Spread Index (ISI)?"
7. "How much does the temperature affect the FFMC?"

Part 2 – Requirements

Considering the data available in the dataset and the questions we wanted to answer, we came to the conclusion that the following relations would be the ones we needed to explore.

List of Requirements

1) Relation (count/avg/..) between fire and location in the natural park by (month/day of the week)
2) Months with the highest fire frequency (histogram)
3) Days of the week with the highest fire frequency (histogram)
4) Relation between burned area (max/min/average) and Meteorological Conditions (Wind Speed/Relative Humidity/Accumulated Precipitation) by months
5) Relation between burned area (max/min/average) and the Canadian Codes (in %): FFMC (Fine Fuel Moisture Code) / DMC (Duff Moisture Code) / DC (Drought Code)
6) Relation between the ISI (Initial Spread Index) and the wind speed
7) Relation between the FFMC and the temperature

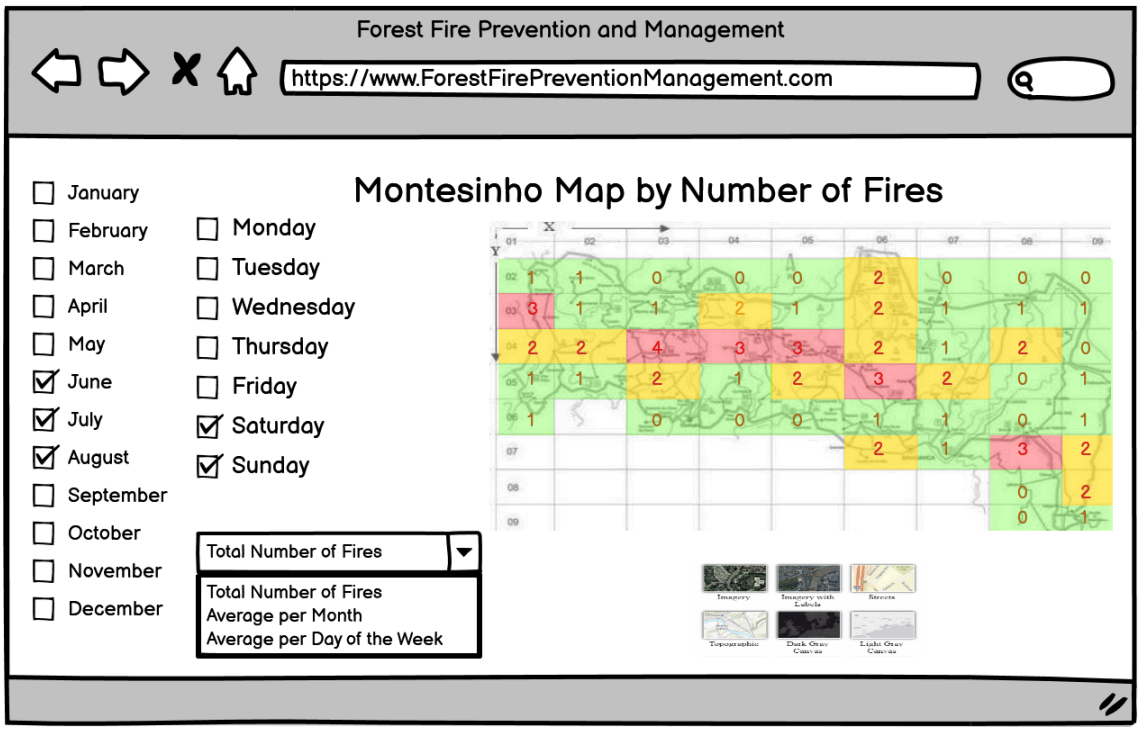

Figure 1: Montesinho Map by Number of Fires

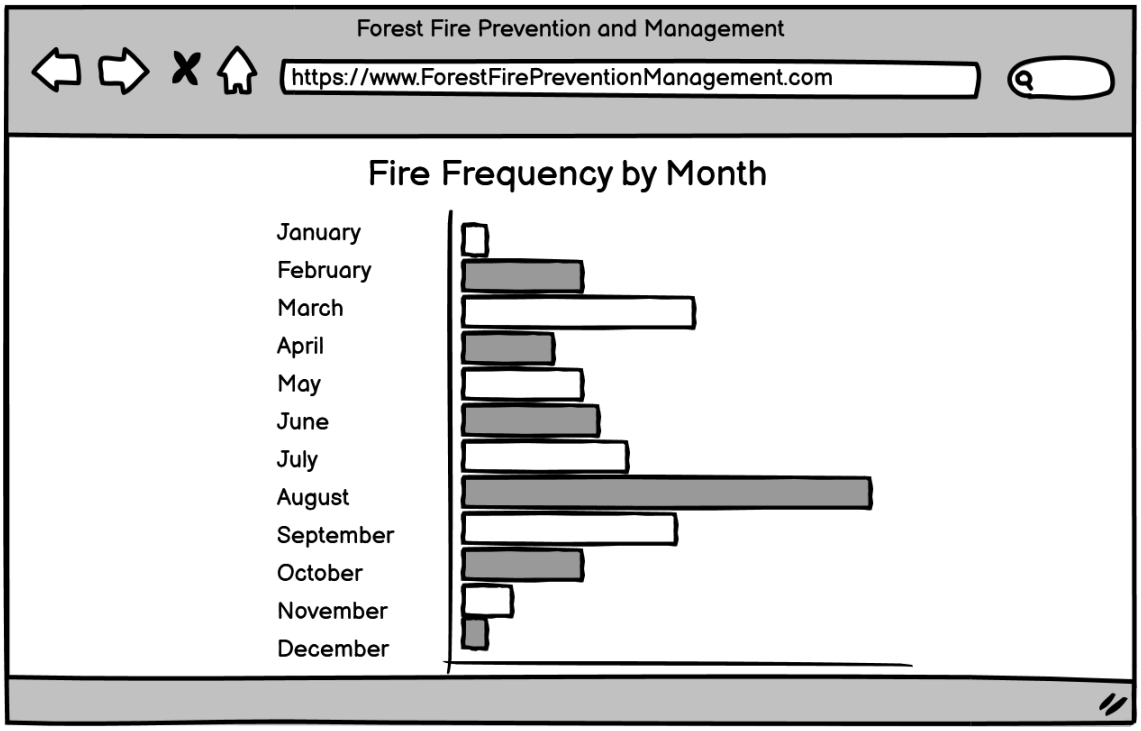

Figure 2: Fire Frequency by Month

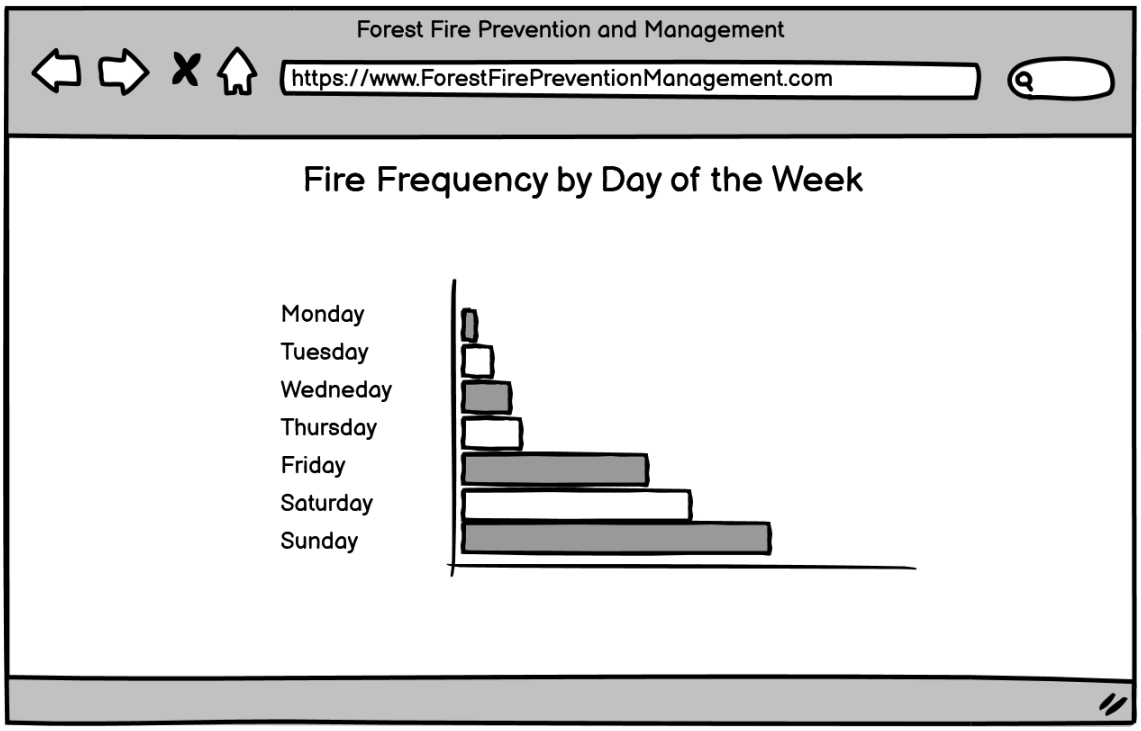

Figure 3: Fire Frequency by Day of the Week

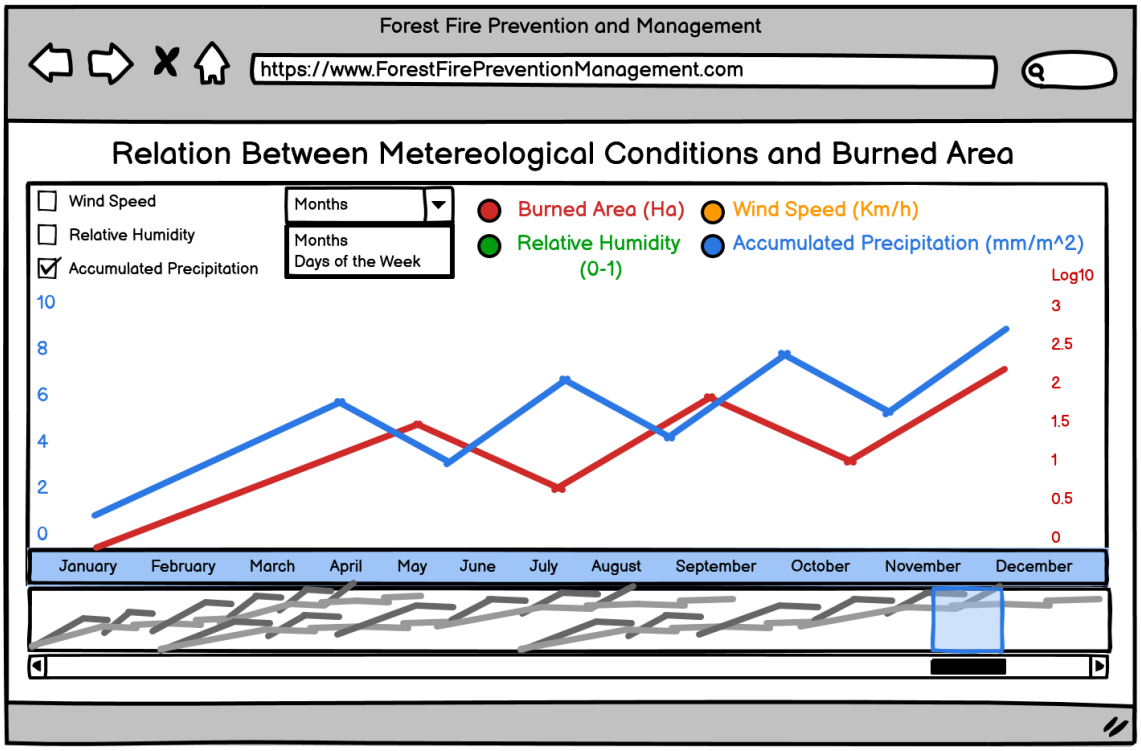

Figure 4: Relation between Meteorological Conditions and Burned Area

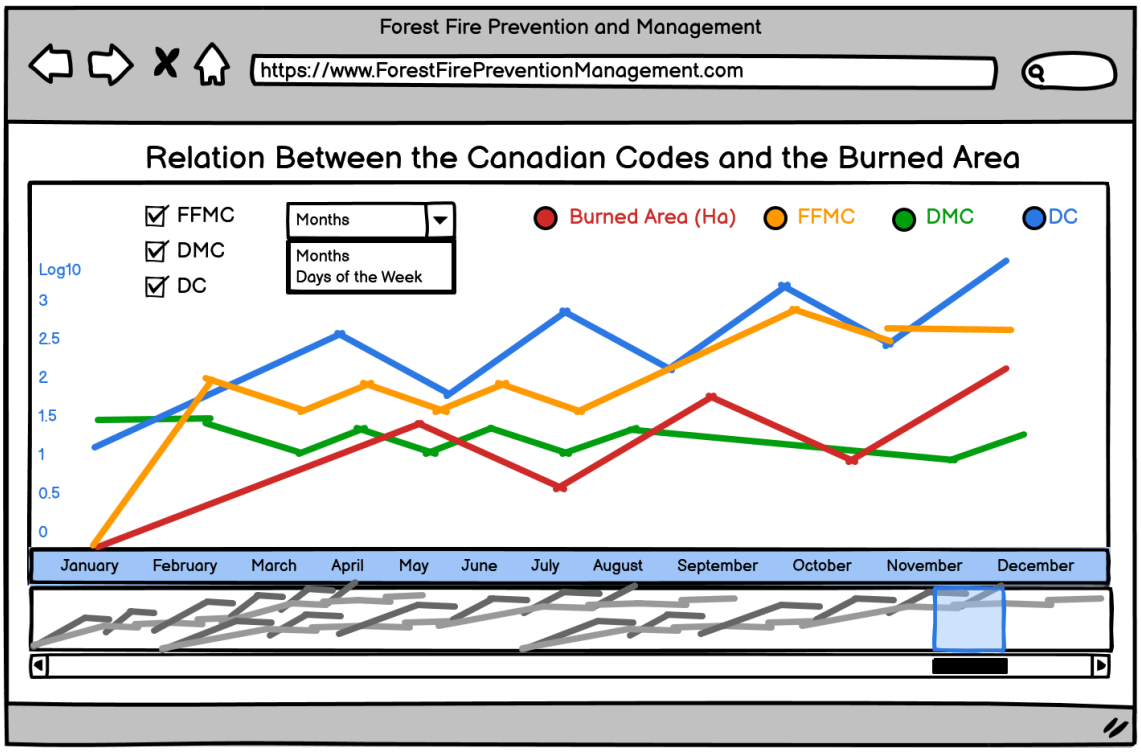

Figure 5: Relation between the Canadian Codes and the Burned Area

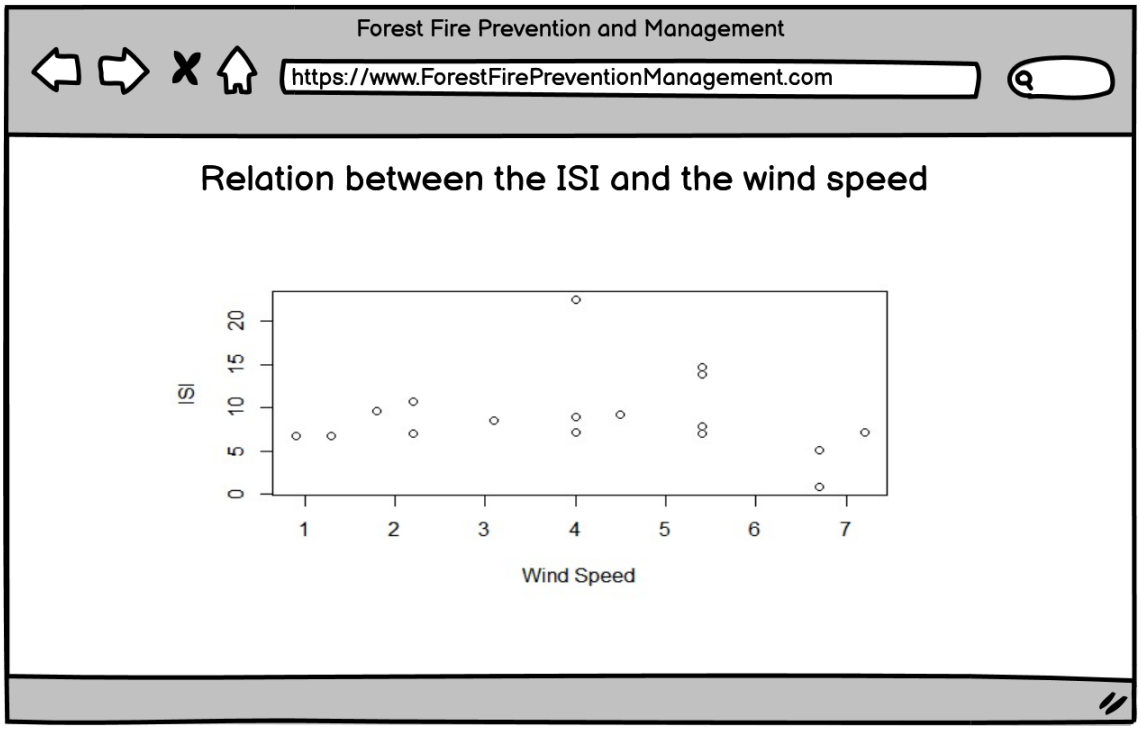

Figure 6: Relation between the ISI and the Wind Speed

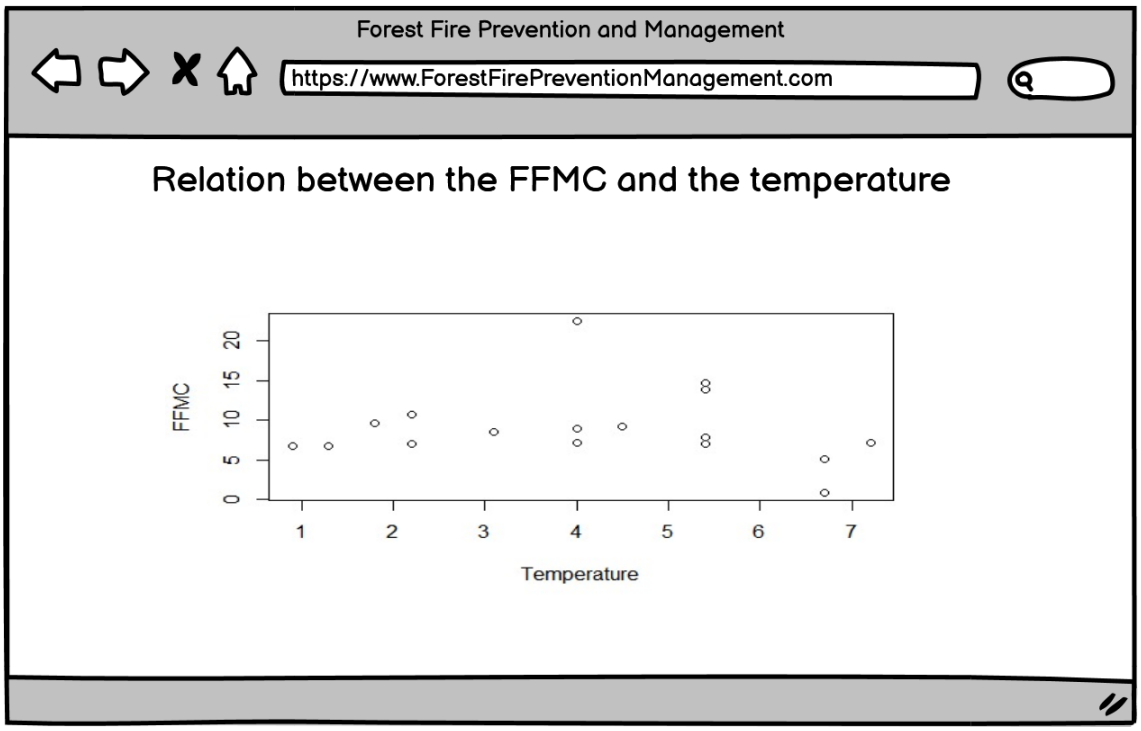

Figure 7: Relation between the FFMC and the Temperature

Part 3 – Design Data Warehouse

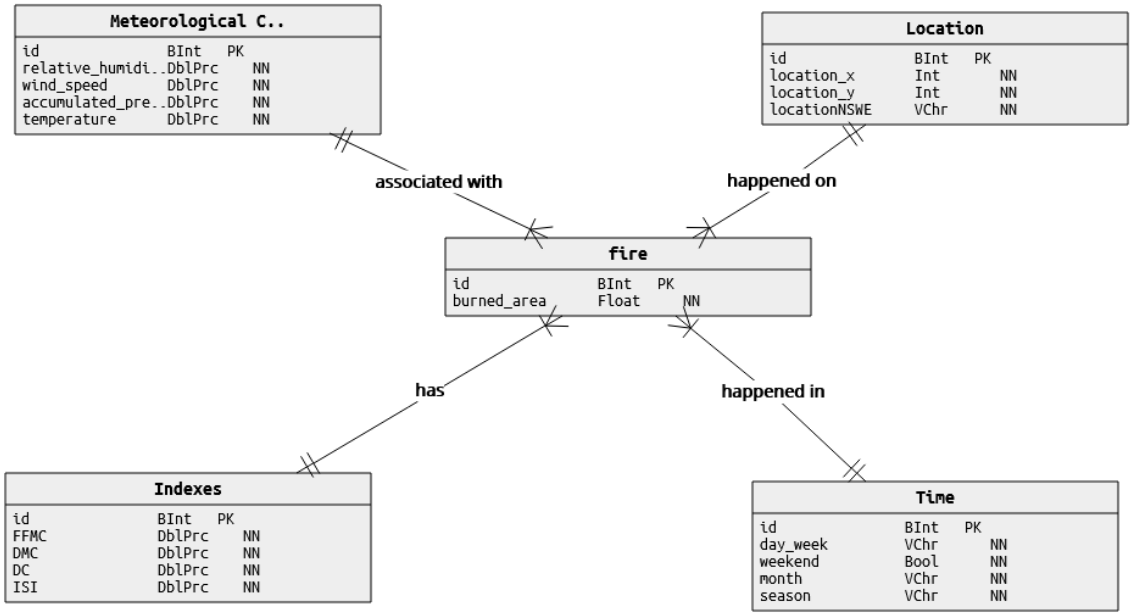

Figure 8: Conceptual Diagram of the Star Scheme

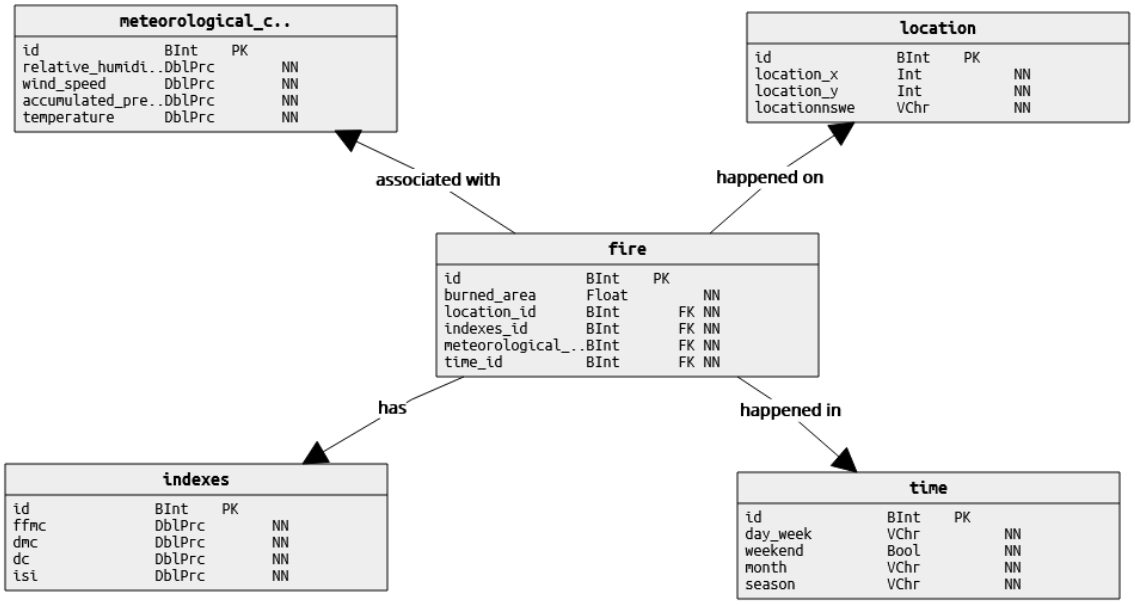

Figure 9: Physical Diagram of the Star Scheme

1. "Stars" in the Model

In our Data Model we decided to have a single star, "fire". Since every record in our dataset is a fire itself, everything else ends up being directly related with the fire (and their burned areas). We can see in the Conceptual Diagram that it has a M..1 relation with all the surrounding dimensions, which is the standard.

2. Facts Table and the facts

As we only have 1 "star", it is natural to have only 1 Facts Table, "fire". In this table we only have 1 fact, since everything else is either the primary key or a foreign key. The fact "burned area" represents the hectares of area burned by the corresponding fire. This fact is considered "additive", since summing burned areas (for example, in a month) is something that has real meaning (total area burned by the fire).

3. Identify the dimensions and their attributes;

We have 4 different dimensions:

1) Meteorological Conditions – This dimension has several attributes related with weather conditions. Notably, we considered Relative Humidity (humidity in %), Wind Speed (km/h), Accumulated Precipitation (30 minutes before the fire, in mm/m^2) and Temperature (in °C).
2) Location – For this dimension we considered the x and y coordinates of the location where each fire took place. These values go from 1 to 9 for x and 2 to 9 for y and are used to easily identify in the Montesinho map the locations of fires. We also added an extra attribute locationNSWE, where we specify the cardinal direction (north, south, east, west, northwest, northeast, southwest, southeast, center).

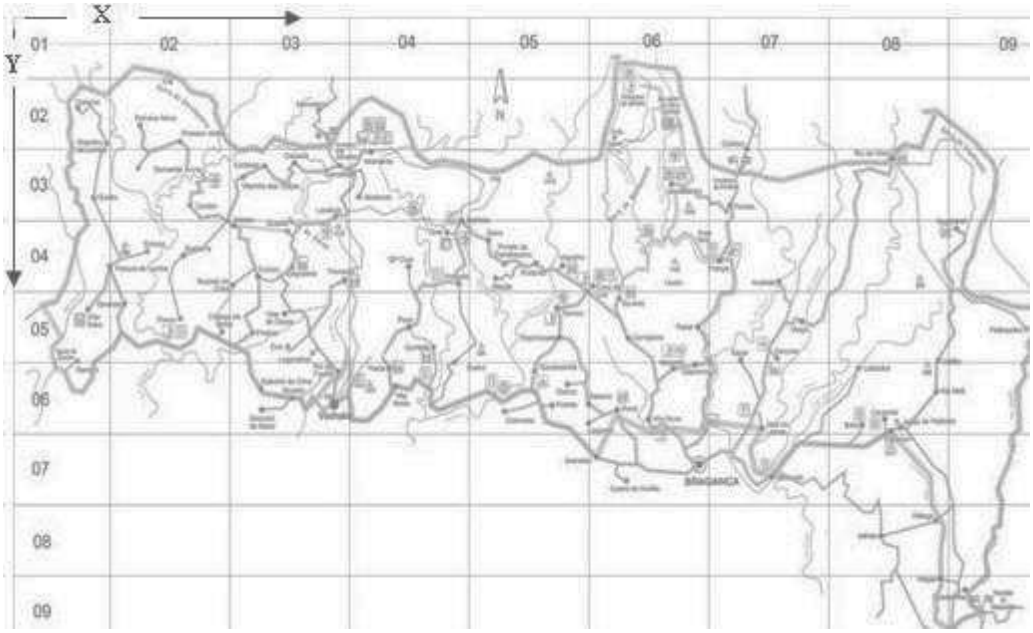

Figure 10: Montesinho Map, as seen in [1]

3) Indexes – This dimension takes into consideration the Canadian Indexes for Fire prevention, which are used to measure the risk of a fire occurrence. These indexes are the FFMC (Fine Fuel Moisture Code), the DMC (Duff Moisture Code), the DC (Drought Code) and the ISI (Initial Spread Index). The FFMC indicates the flammability of the fuel and how easily it can ignite, the DMC indicates the humidity of the layers, the DC

indicates the effects of drought on the forest fuels and the ISI indicates the expected rate of fire spread (based on the wind) **[3]**.

4) Time – In this dimension we consider the different ways of looking at time, the day of the week and the month. We also have attributes to represent the season (summer, autumn, winter, spring) and if a fire happened on a weekend.

4. Define the granularity of the facts.

For the burned area fact, we can observe how much area is burned, with which level of meteorological conditions, on which location, with which level of the Canadian Indexes of Fire Prevention, on which day of the week or month.

Here we decided to allow for higher granularity to be possible for the location and for the time.

In the location's case, we did this by creating an attribute for the cardinal direction, locationNSWE (north, south, east, west, northwest, northeast, southwest, southeast, center), which is useful when the user needs to see information regarding a particular part of the park, without selecting all the specific location points (with x and y).

As for the time, we added attributes to represent (1) the season (summer, autumn, winter, spring), as a higher granularity for the month, since certain seasons will have different impacts on the occurrence of fires and (2) to represent whether a fire happened on a weekend, as a higher granularity of the day of the week, which is helpful since it relates with the occurrence of fires created by man.

Part 4 – Software

We used the tools presented as a basis and then searched for other alternatives that seemed interesting in the web. We also took a close look at the main websites for each software we were analysing, as they all had sections that described their features in detail, as well as their type of license.

The type of license will be a key factor in choosing the software, as well as our own familiarity with it.

Database Server Selection

Since the size of our data source is small, the Data Warehouse will be small as well. As such, size requirements are not a big concern.

Software	Type of License	Database Type	Familiarity
PostgreSQL	Free and Open-Source	Relational	Yes
MySQL	General Public License and Proprietary License	Relational	Yes
Oracle	Proprietary	Relational	No
MongoDB	Free Limited Version	NoSQL	So, so

Table 1: Database Server

Out of these, we decided to go with PostgreSQL, since it has the key factors we are looking for, although MySQL would also be a great choice for the same reasons. In general, PostgreSQL seems to be a common pick for ETL processes, since it heavily reduces operation costs **[4]**.

We picked it over MongoDB, since we ultimately are more experienced in PostgreSQL. That being said, choosing a NoSQL database like MongoDB could be an interesting idea.

ETL and OLAP Software Selection

We decided to represent these 2 choices in 1 table, since ETL and OLAP frequently end up being present in the same tools. This way we can easily visualize which ones have the most features, while also considering which ones are the best for each part. The objective is to understand which ones will be truly necessary.

Software	Type of License	ETL	Reporting/ Dashboards	Familiarity
Pentaho	Community (Open Source)	Yes	Yes	Yes
Talend Open Studio	Free (Open Source), Premium	Yes	Yes	Yes
Domo	Trial, Premium	Yes	Yes	No
Tableau	Public Version	No	Yes	Yes

Table 2: ETL and OLAP

For the ETL part we decided to use Pentaho. Not only are we pretty familiar with it, but it also provides a very intuitive interface, that simplifies the whole process.

To do OLAP we chose Tableau. Tableau allows the creation of very clean dashboards and reports. While comparing it to other choices, we found it to be the best in this department. Creating these is also relatively simple in comparison with the other tools.

Part 5 – ETL

1. Data Sources (Identification and Description)

As mentioned previously, we will use the data we found in **[1]**. This data source was in a .csv file, which we converted to an .xlsx format, which will be used. This data source refers to 517 fires that happened in the Montesinho Natural Park from January 2000 to December 2003.

The data source has the following columns, described in more detail in section 3.3:

X (location)	Y (location)
month	day
FFMC	DMC
DC	ISI
Temp (Temperature)	RH (Relative Humidity)
wind	Rain (Accumulated Precipitation)
area	

2. Overall ETL plan

First of all, in order to visualize the Relative Humidity along with the other meteorological conditions, we needed to change it from percentage (1-100%) to values from 0 to 1, which can be done with a simple division by 100.

As for the burned Area, using a logarithmic scale makes sense, since a lot of its values are close to 0, but there are also enormous values, close to 1000, as explained in **[5]**. However, for values between 0 and 1 (excluding 0 and 1), this would give negative values. As a solution to this, we will first apply a translation of 1, as suggested in **[6]**. So the final transformation will be ln(value + 1).

A logarithmic scale was also be used for the Canadian Codes (FFMC, DMC and DC). This is necessary, as we want to visualize them together, but the DC

tends to have much bigger values (goes up to 860.6, while the others don't even reach 300).

There are also attributes we decided to create to add the possibility of higher granularity, as explained before. This meant we needed to add attributes for (1) the locationNSWE, (2) the weekend and (3) the season, which are based on (1) the x and y, (2) the day of the week and (3) the month.

By inspecting the dataset we noticed that there weren't any null values, so we realized we didn't need to do this type of cleaning (value correction).

3-4. Staging Area Description and Explanation of the actions in the plan

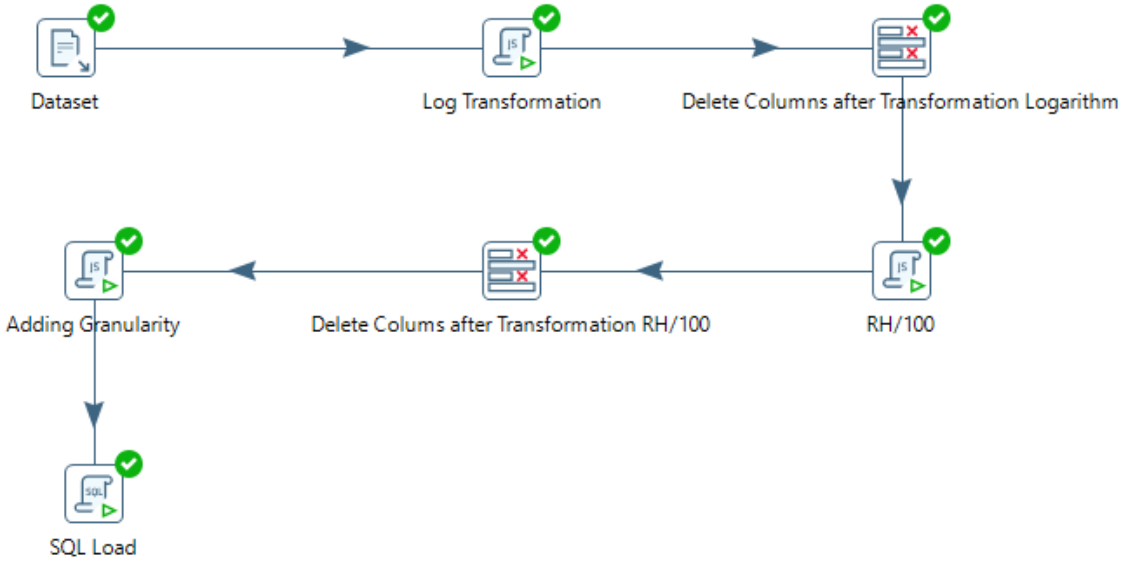

Figure 11: Staging Area

Extract

First, we use a "CSV file input" to read the dataset, while making sure to use the correct separator.

Transform

Then, in "Log Transformation", we apply the necessary ln(value+1) transformations mentioned in 5.2. This is done with a "Modified Javacript Value", which was chosen due to the lack of a "ln" function in Pentaho's commonly used "Calculator".

Afterwards we use "Delete Columns after Transformation Logarithm", which is a "Select Values", in order to remove the original 4 columns that were used to create their ln(value+1) counterparts, as they are no longer needed.

Another "Modified Javacript Value" is used right after. This one, "RH/100", simply converts RH to values from 0 to 1, as mentioned in 5.2.

"Delete Colums after Transformation RH/100" simply removes the original RH column.

In the step "Adding Granularity", we use a "Modified Javacript Value" in order to create the new attributes. We essentially use "if clauses" to map (1) x and y coordinates to cardinal directions (locationNSWE), (2) days of the week into a Boolean attribute that defines whether a day is part of a weekend (we decided to count Fridays as well) and (3) months into seasons (summer, autumn, winter and spring).

Load

Finally, we use a "Execute SQL Script", since it's necessary to insert the data in the database. For these queries, we had to make sure we weren't repeating values (for example, putting Monday-July multiple times), as it would waste space. As such, we made adjustments (as seen in the next figure) to make sure we were only inserting rows that weren't there already.

```
INSERT INTO location (location_x, location_y, locationnswe)
    SELECT ?, ?, '?'
    WHERE NOT EXISTS (SELECT *
                      FROM location
                      WHERE location_x = ? and location_y = ? and locationnswe = '?');

INSERT INTO time (day_week, month, season, weekend)
    SELECT '?', '?', '?', '?'
    WHERE NOT EXISTS (SELECT *
                      FROM time
                      WHERE day_week = '?' and month = '?' and season = '?' and weekend = '?');

INSERT INTO meteorological_conditions (relative_humidity, wind_speed, accumulated_precipitation,
                                       temperature)
    SELECT ?, ?, ?, ?
    WHERE NOT EXISTS (SELECT *
                      FROM meteorological_conditions
                      WHERE relative_humidity = ? and wind_speed = ? and
                            accumulated_precipitation = ? and temperature = ?);

INSERT INTO indexes (ffmc, dmc, dc, isi)
    SELECT ?, ?, ?, ?
    WHERE NOT EXISTS (SELECT *
                      FROM indexes
                      WHERE ffmc = ? and dmc = ? and dc = ? and isi = ?);

INSERT INTO fire (burned_area, location_id, indexes_id, meteorological_conditions_id, time_id)
VALUES (?,
           (SELECT id
           FROM location
           WHERE location_x = ? and location_y = ?),
               (SELECT id
               FROM indexes
               WHERE ffmc = ? and dmc = ? and dc = ? and isi = ?),
                   (SELECT id
                   FROM meteorological_conditions
                   WHERE relative_humidity = ?
                           and wind_speed = ? and accumulated_precipitation = ? and temperature = ?),
                       (SELECT id
                       FROM time
                       WHERE day_week = '?' and month = '?')
       )
```

Figure 12: SQL Insert statements for all the tables

5. Major Challenges of the implementation

"Modified Javacript Value" were picked over "User Defined Java Expression" since we had problems with the default output from java, which resulted in numbers separated by "," instead of ".". This also happened with Javascript, but correcting it here was far easier, as all it took was replacing the type of the output with "BigNumber". Another solution we tried was changing the location that Pentaho associates with the user (which could be influencing the type of decimal separator). However, this did not work as expected.

While dividing RH by 100 we couldn't find a way to divide every number by a constant instead of a column, which is why we resorted to using a "Modified Javacript Value" instead.

We had numerous problems inserting the data into the datatbase. First we tried using the "Table Output" we experimented with in the tutorials, but it was doing a lot of unwanted changes in terms of the tables themselves (deleting tables it shouldn't, deleting tables so it could create them again, etc). On top that it was having problems inserting the data correctly, resulting in a lot of errors that we couldn't really understand. We also tried to use "Insert/update", but ran into similar issues.

To solve that issue, we ended up using "Execute SQL Script", which gave us freedom in the creation of the queries.

We also had some issues with the number of decimal cases some numbers had. Notably for the meteorological conditions and the indexes, which were originally floats in the database, but had to be changed to doubles to match the SELECT's done in **figure 13**.

6. Metrics

6.1 – Source Data Size (.csv): 25KB.

6.2 – Size of Data on the First Load (including indexes):

1) Fire – 96KB
2) Indexes – 56KB
3) Location – 32KB
4) Meteorological Conditions – 88KB
5) Time – 32 KB

The size difference in relation to the source is very likely due to the differences between the .csv and the PostgreSQL formats (otherwise the SQL should have a smaller size). The total size is **304 KB**.

6.3 – Time spent on the First Load:

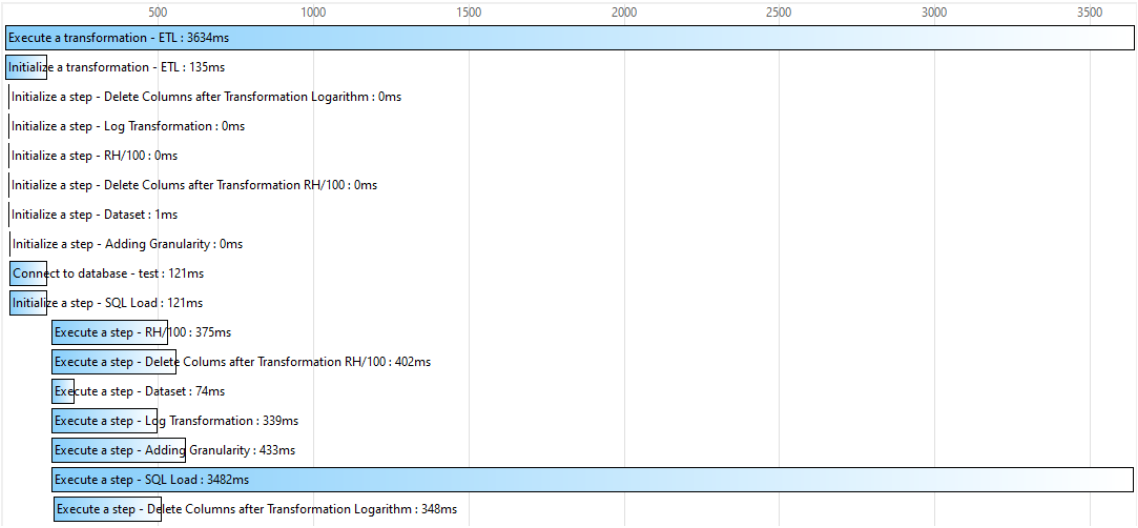

Figure 13: Time Spent on the First Load

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	Dataset	0	0	517	518	0	0	0	0	Finished	0.1s	6,475
2	Log Transformation	0	517	517	0	0	0	0	0	Finished	0.3s	1,503
3	Delete Columns after Transformation Logarithm	0	517	517	0	0	0	0	0	Finished	0.4s	1,432
4	RH/100	0	517	517	0	0	0	0	0	Finished	0.4s	1,360
5	Delete Colums after Transformation RH/100	0	517	517	0	0	0	0	0	Finished	0.4s	1,270
6	Adding Granularity	0	517	517	0	0	0	0	0	Finished	0.4s	1,183
7	SQL Load	0	517	517	0	0	0	0	0	Finished	3.5s	148

Figure 14: Information about the First Load (Time, Speed)

7. Problems with the Source and Data Warehouse Data

The only problem we couldn't solve was the lack of information regarding the year when each fire occurred. This seems impossible to fix, as there is no such information in the source and the dates are presented in a seemingly random order.

8. Strategy for future updates regarding the ETL process

In our scenario, since we are using data from a relatively old study that lasted 3 years, we are not expecting a constant stream of updates. Assuming that were the case, a solution could be to add a Kafka Consumer to receive new data and automate the Extract part of the ETL process. In that case, we should update once per hour. We consider this to be a good idea since we want fires registered

one day to become available in the database that same day. It's worth noting that in our case there isn't a concern of an excessive amount of data.

If a relatively similar study were to be performed, also lasting a couple of years, we could implement the strategy above (assuming we had access to said data).

9 - 10. Presentation of the OLAP Data and Analysis Performed

Figure 15: Montesinho Map by Number of Fires

In our first dashboard we present the number of fires by location. Each location has a number and colour associated with it, which are the number of fires that happened on that location and how that number compares with the other locations (going from green to red). The user can also hover on any location to get clear information about the number of fires and location.

It is also possible to see the average number of fires for the period of time selected, as seen in the top left corner.

The end user can change the selection of data presented through checkboxes. Notably, it is possible to select the months and days of the week that are being considered. This gives freedom to the user in terms of the specific temporal context they want to analyse. We also made sure to allow the user to select if they only wanted to see the weekends or specific seasons, to give a higher level of granularity, as a way to facilitate analysis of specific contexts.

To add on to that, we also have a checkbox ("Cardinal Direction") that allows users to select specific cardinal directions for the locations. So, for example, a user can choose to only see the fires that happened on the northwest or the south.

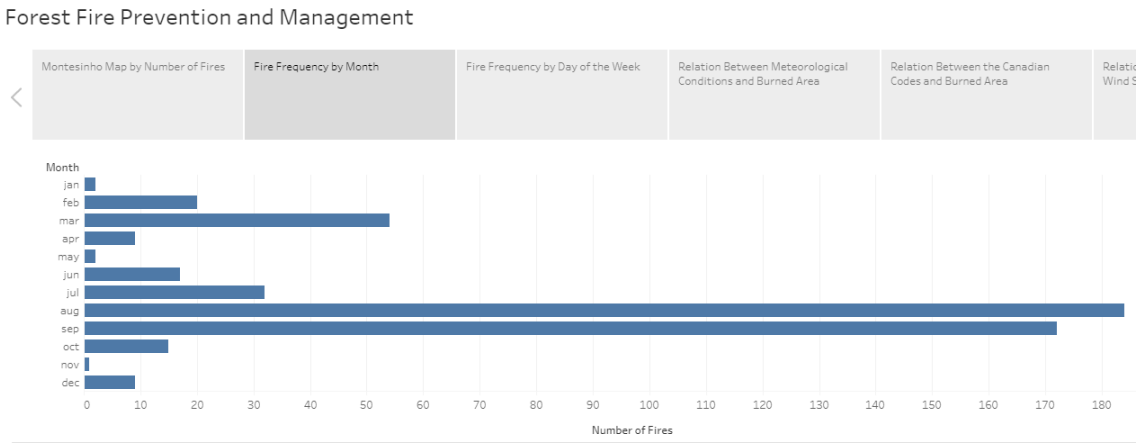

Figure 16: Fire Frequency by Month

We allow users to easily switch between dashboards, by simply clicking on the one they want to see on the slider shown above. This is true for all of them.

In this report we can see a graph that shows the number of fires that happened each month. With this, it should be possible to draw conclusions regarding the likeliness of fires occurring in a certain month.

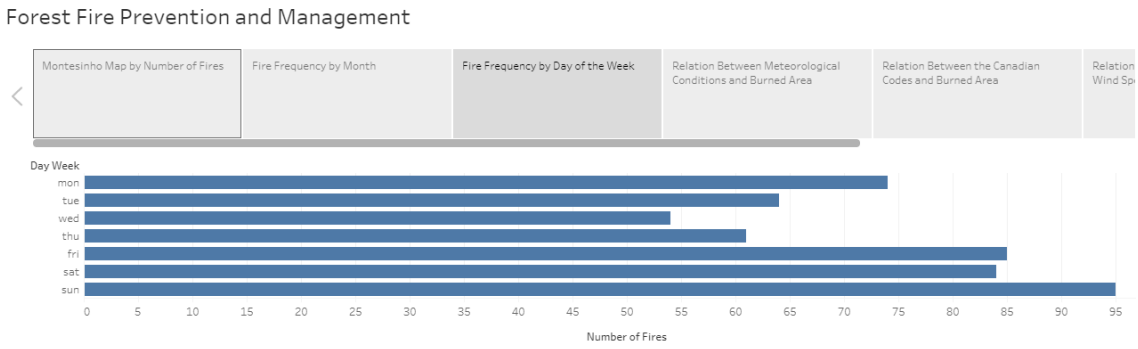

Figure 17: Fire Frequency by Day of the Week

This report is similar to the previous one, but shows the number of fires per day of the week instead. In this case, the idea is to try to see if there are a great number of fires caused by human hand (which would be related with higher probabilities on the weekends).

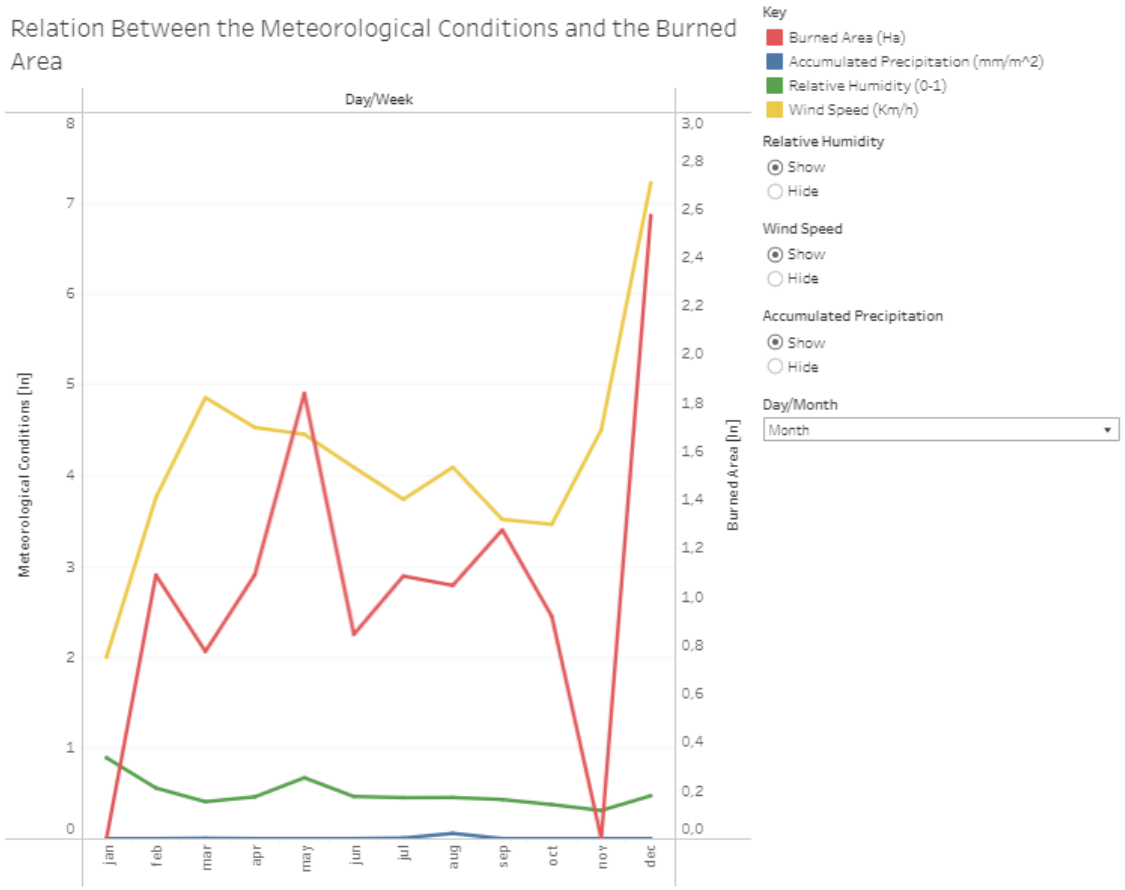

Figure 18: Relation between Meteorological Conditions and Burned Area

In this dashboard, a user can see a graph that shows the Burned Area and the Meteorological Conditions. It is thus possible to analyse the relation between

the area that was burned and the weather conditions. This can be shown either by month or by day of the week, which can be changed by the user with a dropdown menu.

On top of that, the user can also decide to either show or hide certain weather conditions. This can allow for cleaner visualizations of the data, in cases where there is a focus on a particular weather condition.

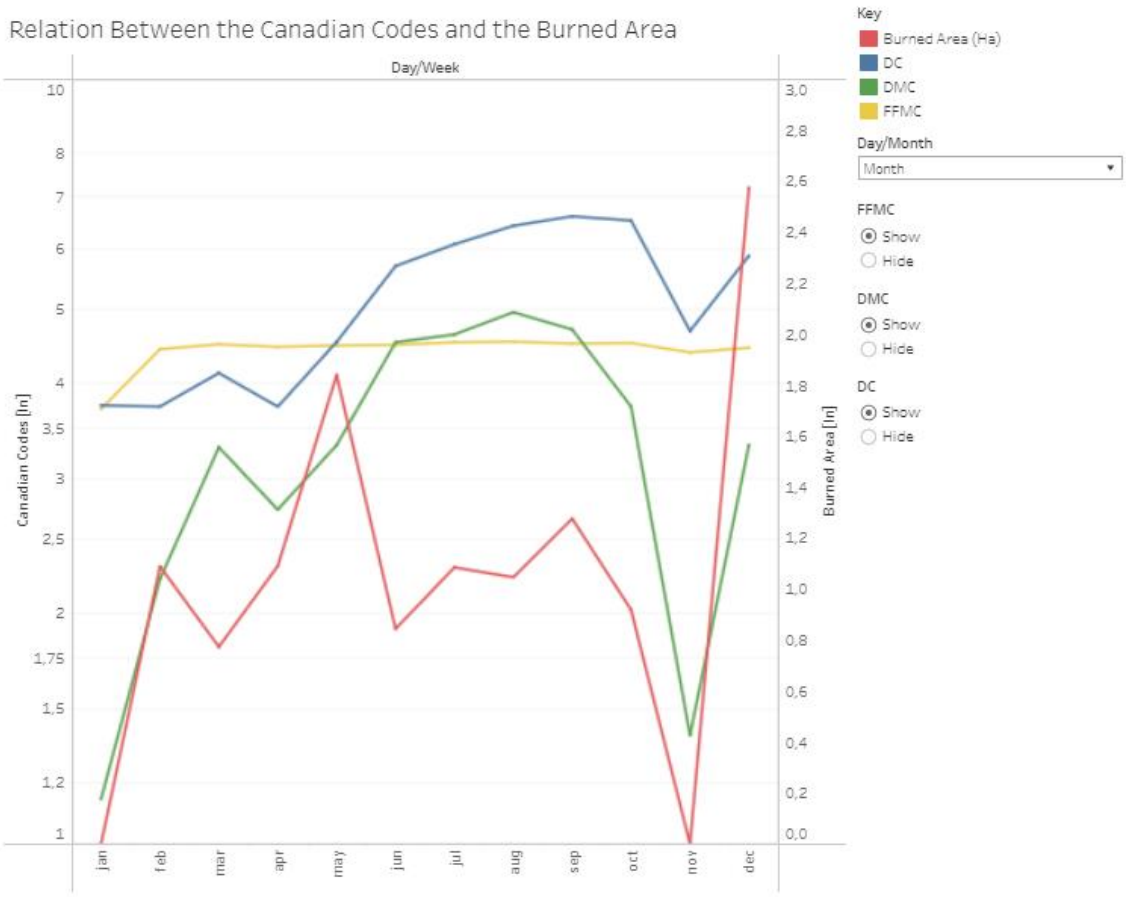

Figure 19: Relation between the Canadian Codes and Burned Area

As for this one, there is a graph showing the Burned Area and the Canadian Indexes of Fire Prevention. Similarly to the last one, the user can see the relations between the area burned by fires and the Canadian Codes. Once again, this can be shown either by month or by day of the week, which can be changed by the user with a dropdown menu.

It is also possible to either show or hide certain indexes, which can be useful for specific purposes.

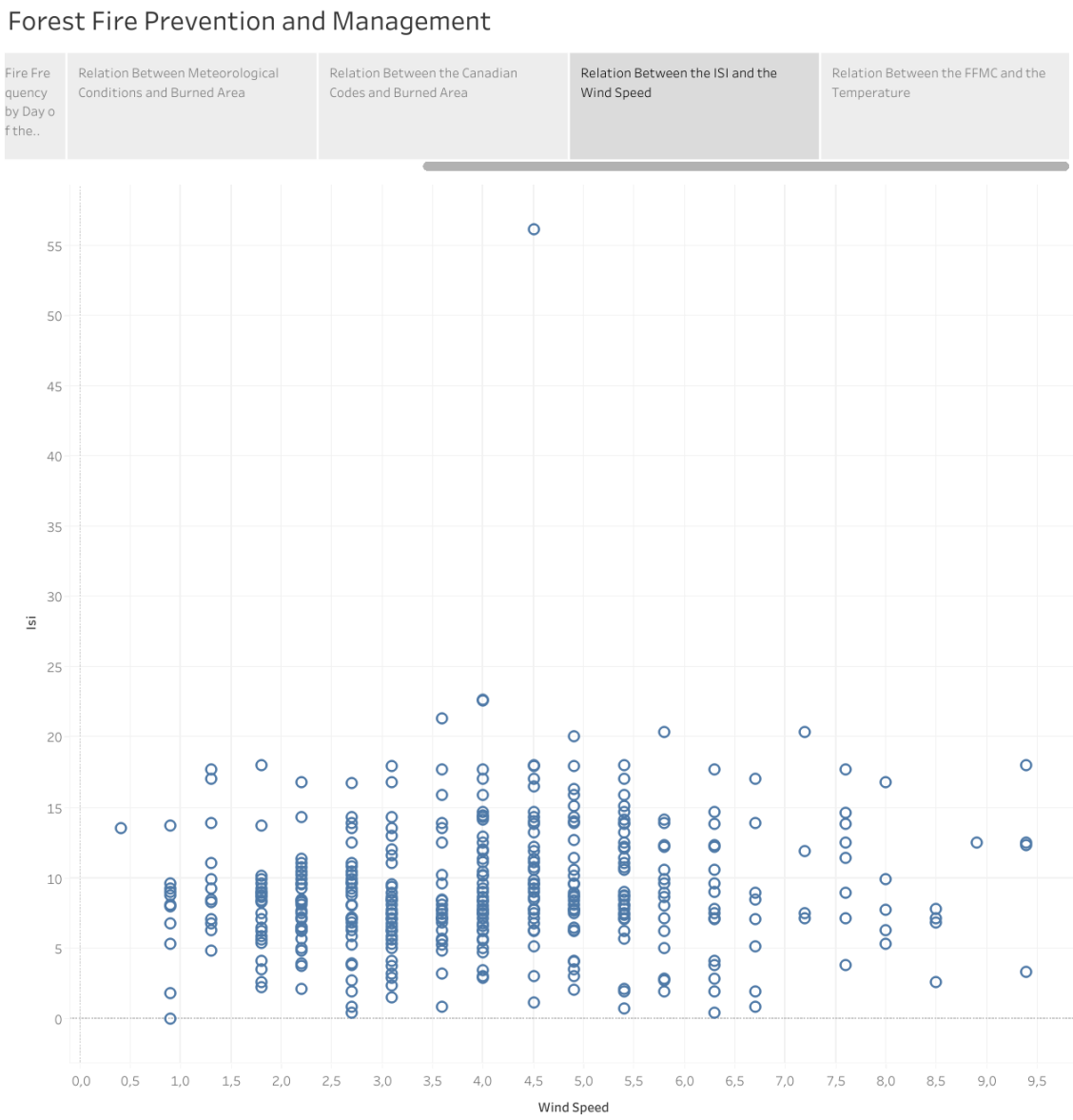

Figure 20: Relation between the ISI and Wind Speed

This report shows the relation between the ISI and the Wind Speed through a scatter plot. This simplifies the process of understanding the relation between the 2, as a way to draw conclusions. In this case, it can be useful to understand how much impact the Wind Speed has on the ISI (Initial Spread Index).

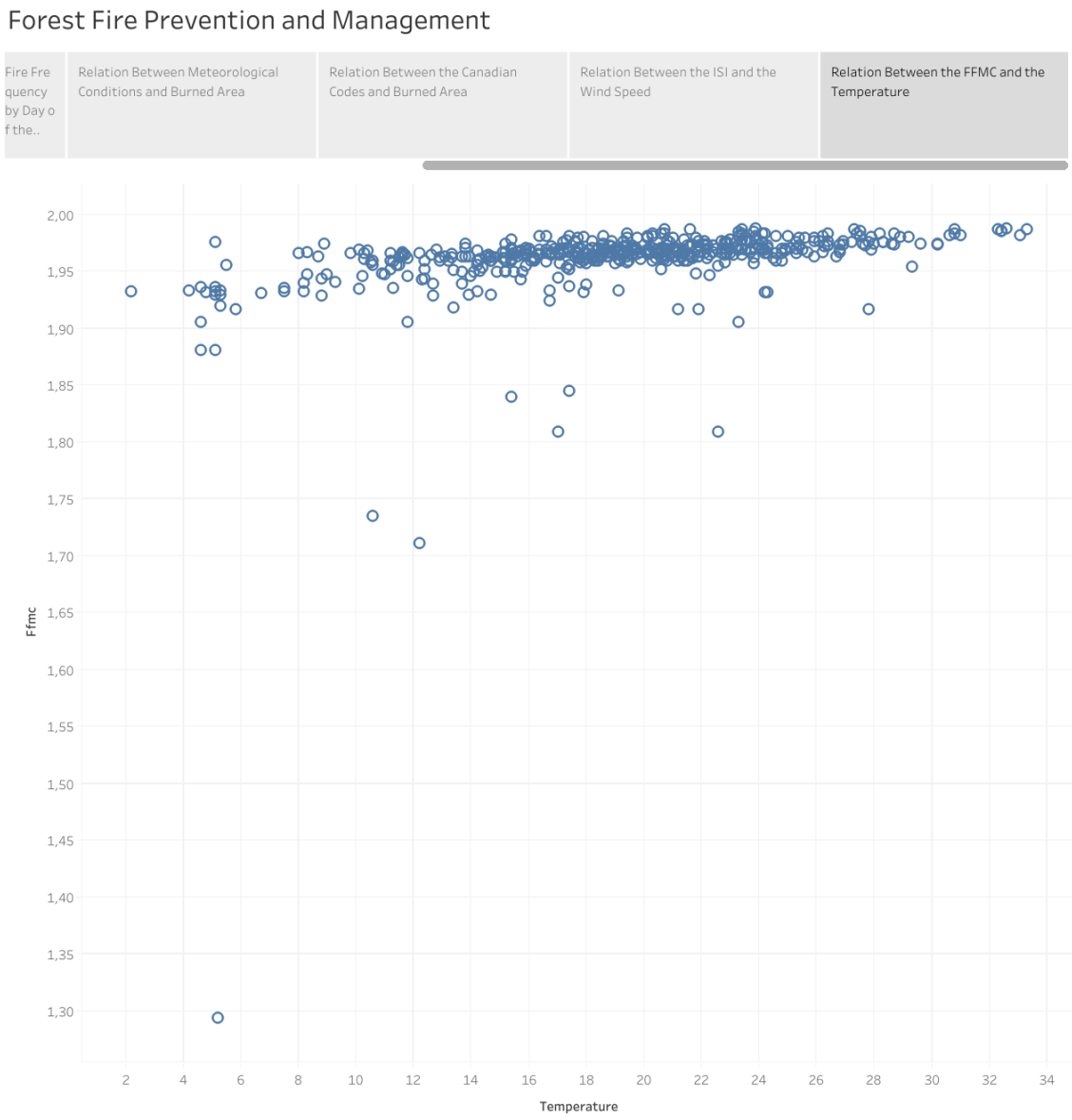

Figure 21: Relation between the FFMC and the Temperature

This report is a scatter plot used to represent the relation between the FFMC and the temperature. It is possible to draw conclusions regarding the effects of high/low temperatures on the FFMC.

11. Presentation of the initial findings and discussion

1 - For the dashboard in figure 15 we noticed that there is a greater concentration of a high number of fires in the center of the park, with the outlier being the location close to the bottom right of the map. We can conclude that areas on the outer bounds of the park are generally safer than the ones inside, as they are less likely to suffer from fire occurrences. This is probably due to the higher concentration of vegetation in the innermost parts of the Montesinho park.

While selecting the weekends using the correct drop down, we realized that in general there is a bigger spread of a high number of fires to the middle left. However, there are also 3 specific spots on the right that have a massive concentration of fires, even more so than the ones on the left. This might mean that the right part of the park has clear locations that are more vulnerable to this type of disaster. We can see this behaviour on the map below, where there are 3 values on the left over 20, while the average is only 8,8.

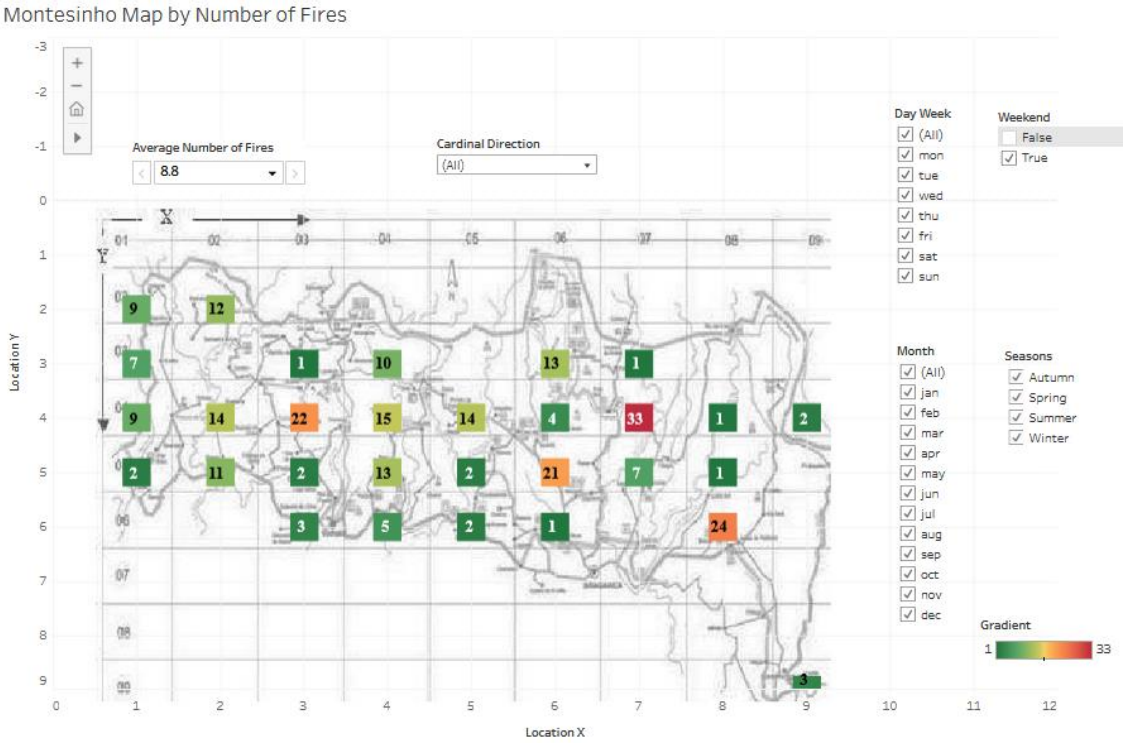

Figure 22: Fires in the Montesinho Park during the weekends

2 – While analysing the histogram in figure 16, we found that August had the highest number of fires. This was expected, as it is the hottest month. Very closely following it we have September, which is very likely this high due to being the month right next to August. In general, it seems like these are the months where fire prevention is the most crucial.

One of our surprises was that June and July, while still quite heavy on the number of fires, were so much lower than August. Meanwhile, March unexpectedly had a higher number, which didn't make too much sense at first. On a second analysis we noticed that it was possibly due to a heavy number of manmade fires during March, relatively to the other months. This could explain the inflated value.

3 – By analysing the histogram in figure 17, we immediately noticed something we had predicted: there is a higher number of fires during weekends. Not only that, but the next highest numbers come from Friday and Monday. This indicates a high possibility of fires caused by man, as weekends give way to more

time for planning and executing these types of disasters. Friday is also high, as it is right before that. Meanwhile Monday suffers from being right after, as there could be fires that spread more easily due to previous ones. In general, it is advisable to be especially cautious during these 4 days of the week.

4 – As for figure 18, we quickly realized that the accumulated precipitation was nearly constant at 0. This makes sense, since it is harder for a fire to spread when there is the possibility of rain. In fact, the only time it happened was on August, likely due to the very high temperatures.

Regarding the wind speed, we noticed that its highest point (in December) was highly correlated with the burned area. By contrast, the Wind Speed in the first peak of the graph (in March) didn't seem to have much impact. This could be due to the fires in December having wind as a primary factor of fire spread. If this is the case, this month requires special caution regarding measures related with the wind.

As for the Relative Humidity, it doesn't seem to have a big connection with the area burned, as their lines follow very different trajectories.

Finally, we found it strange that December had the highest burned area. However, by analysing the source data we noticed that the fires in December, while not having massive amounts of burned area, all burned a decent amount of it. Meanwhile the other months tended to have multiple fires that didn't burn even 1 Ha.

5 – In the dashboard from figure 19, we can observe that the lines representing the DMC, the DC and the burned area quite similar in terms of behaviour. This shows that these indexes had a very large impact on the amount of area burned. We believe that paying close attention to them in particular could be very powerful in terms of fire prevention.

As for the FFMC, it remained nearly constant, so it doesn't seem to have had a big impact regarding the data of the study.

6 – On the top middle of the graph in figure 20, we can spot a pretty massive outlier, having more than double the ISI of any other. It seems like this might be caused by something other than the wind speed itself.

As for the rest of the graph, we couldn't really conclude much, since the points are all very spread-out, with seemingly no real pattern.

7 – Meanwhile the graph from figure 21 has a really low value compared to everything else, on the bottom left. Again, this could be due to other factors not related to the ones explored on the graph.

Looking at the graph, we can notice that there is a slight tendency for the FFMC to increase as the temperature rises, as seen in the gradually higher concentration of points above. However, this does not seem to be a massive correlation, as most of the points tend to be quite high, almost forming a constant line that gets slowly but surely higher. We can access that a decent amount of attention should be paid to the influence of the temperature on the FFMC, but it shouldn't be overstated.

12. The results seen from a business perspective

As we explained, there is a high probability of manmade fire during the weekends and the 2 days surrounding them (Monday and Friday). As a result, we advise the correct deployment of GNR officers to combat this issue, specifically on those days. This could make it easier to catch the people starting these fires, as there would be a previous preparation, which could in turn reduce these occurrences. The same could be applied to March, which seems to also be a target of these acts.

Since August and September have such an enormous probability of fire, the ICNF, which has several plans and protocols in place, should adjust them accordingly, as to take into account the massive risks associated with these months.

Regarding the locations where the fires took place, we can advise extra caution regarding the middle area of the park, as it is a more vulnerable target for these disasters.

13. Performance Optimization

In order to more quickly find fires from (1) certain seasons, (2) during weekends and (3) in certain cardinal locations, we made sure to add granularity to the "time" dimension for points 1 and 2 and to the "location" dimension for point 3.

As for indexes, we used Foreign Keys for each of the dimensions of the Star Scheme, as means of optimizing performance.

It is also worth noticing that we made sure to remove repeated combinations of values (for day-week, indexes, meteorological conditions, location). This was done to avoid having overly redundant data on the table, which ends up benefitting performance, as useless rows won't have to be read.

We ultimately chose not to use aggregates (materialized views). However, there are some places where they could have been useful. For example, an aggregate with the mean in case the user chooses to see the fires from a weekend and vice-versa.

Part 6 – Machine Learning/Data Mining

6.a. Classification Study

6.a.1 Objective

Our objective is to create a model that can **classify** whether there will be **burned area** or not, for new fires. With this, it should be possible to identify the most threatening ones. This information can then be used to dispatch firemen accordingly.

6.a.2 Source Data Model and Sets of Attributes used

Source Data Model - The source data model will be the same one obtained at the end of the ETL process, as depicted previously in figure **8**. We exported the tables from the PostgreSQL database as **csv** files, which are then read in Python.

Unused Attributes - Since the **cardinal direction**, the **weekend** and the **season** were added in a previous part of the project, with a completely different goal (speeding up queries) they will **not be used** in the machine learning techniques.

Sets of Attributes Used - In terms of the attributes selected for this, we decided to do something **similar** to **[5]**, where different attribute selection setups are compared. In our case we decided to compare these specific setups:

1. LT (location and time)
2. LTM (location, time and meteorology)
3. LTCM (location, time, Canadian indexes and meteorology)
4. LCM (location, Canadian indexes and meteorology)
5. M (meteorology)

3 of these are different from any setup used in that study, while 2 of them are the same. The idea is to try different setups to see if we can find some new interesting conclusions, as well as verify the previously obtained results. This is even more relevant for the second objective, where we use regressions, which were the type of machine learning techniques used in that study.

To reiterate, the mapping of dimensions-attributes is the following (excluding the unused attributes).

Indexes – **FFMC**, **DMC**, **DC**, **ISI**

Location – **X**, **Y**

Meteorology – **relative humidity**, **wind speed**, **accumulated precipitation**, **temperature**

Time - **day of the week**, **month**

6.a.3 Data Preparation Activities

Pre-processing 1 - Fixing the issue of having fewer rows in the dimensions - Since we removed duplicates in the dimensions during ETL (e.g. repeated combinations of day of the week and month), we first need to match the ids from each dimension to the ids in the fireDataframe, to create new dataframes with the same number of rows as the fireDataframe.

In the following image, we can see the creation of the new dataframes, by following that strategy.

```
for i in range(length):
    position = fireDataframe.indexes_id[i] - 1
    ffmc = indexesDataframe.ffmc[position]
    dmc = indexesDataframe.dmc[position]
    dc = indexesDataframe.dc[position]
    isi = indexesDataframe.isi[position]
    new_row = {'ffmc':ffmc, 'dmc':dmc, 'dc':dc, 'isi':isi}
    indexesDataframe2 = indexesDataframe2.append(new_row, ignore_index=True)

    position = fireDataframe.location_id[i] - 1
    x = locationDataframe.X[position]
    y = locationDataframe.Y[position]
    new_row = {'X':x, 'Y':y}
    locationDataframe2 = locationDataframe2.append(new_row, ignore_index=True)

    position = fireDataframe.meteorological_conditions_id[i] - 1
    relative_humidity = meteorologyDataframe.relative_humidity[position]
    wind_speed = meteorologyDataframe.wind_speed[position]
    accumulated_precipitation = meteorologyDataframe.accumulated_precipitation[position]
    temperature = meteorologyDataframe.temperature[position]
    new_row = {'relative_humidity':relative_humidity, 'wind_speed':wind_speed, 'accumulat
    meteorologyDataframe2 = meteorologyDataframe2.append(new_row, ignore_index=True)

    position = fireDataframe.time_id[i] - 1
    day_week = timeDataframe.day_week[position]
    month = timeDataframe.month[position]
    new_row = {'day_week':day_week, 'month':month}
    timeDataframe2 = timeDataframe2.append(new_row, ignore_index=True)
```

Figure 12: Creation of the new dataframes, all with the same number of rows as the fireDataframe

Pre-processing 2 - Convert 'day_week' and 'month' so they are usable – Here we simply mapped those 2 attributes into numbers, going from 0 to 6 for days of the week and 0 to 11 for months. This makes sense, since they both have a clear order (ordinal). This is necessary to later be able to fit the models.

```
timeDataframe.loc[timeDataframe["day_week"] == "mon", "day_week"] = 0
timeDataframe.loc[timeDataframe["day_week"] == "tue", "day_week"] = 1
timeDataframe.loc[timeDataframe["day_week"] == "wed", "day_week"] = 2
```

Figure 13: Example of the mapping that was done

Pre-processing 3 - Inverse of ln(x+1) – During the ETL process we had used an ln(x+1) transformation for visualization purposes (for the burned area, ffmc, dmc and dc). These no longer make sense in this context, so we want to invert them.

Also, the tests we did **without this pre-processing** gave **worse results** in terms of the model scores.

To accomplish this, we use "exp(value) –1". Then we turn any negatives into 0, as suggested in **[5]**.

```
#Pre-processing 3 - Inverse of ln(x+1)

indexesDataframe.ffmc = numpy.exp(indexesDataframe.ffmc) - 1
indexesDataframe.ffmc = indexesDataframe.ffmc.clip(lower=0) #If negative -> 0
```

Figure 14: Example of the inversion

Pre-processing 4 - Create dataframe that will be used for classification – For this Classification Study we decided to classify whether there will be burned area or not. For this purpose, we created a new dataframe, with a column that represents the existence of burned area.

```
booleanBurnedAreaDataframe.loc[booleanBurnedAreaDataframe["burned_area"] > 0, "burned_area"] = 1
booleanBurnedAreaDataframe.loc[booleanBurnedAreaDataframe["burned_area"] == 0, "burned_area"] = 0
```

Figure 15: Creation of the new dataframe

Pre-processing 5 - Separate array into input and output components – For this part we made sure to create the different sets of inputs mentioned before in 6.a.2.

```
#Pre-processing 5 - Separate array into input and output components

LT = numpy.concatenate( (location[:,0:2], time[:,0:2]), axis=1) #location and time
LTM = numpy.concatenate( (location[:,0:2], time[:,0:2], meteorology[:,0:4]), axis=1) #location, time
LTCM = numpy.concatenate( (location[:,0:2], time[:,0:2], indexes[:,0:4], meteorology[:,0:4]), axis=1)
LCM = numpy.concatenate( (location[:,0:2], indexes[:,0:4], meteorology[:,0:4]), axis=1) #location, Ca
M = meteorology[:,0:4] #meteorology

Y = booleanBurnedArea[:,0]
```

Figure 16: Creation of the input and output components

Pre-processing 6 - Standardize the data – We tried using (1) Min-max scaling standardization and (2) a typical standardization with mean 0 and standard deviation of 1. Ultimately we decided to go with the second option, since that type of standardization is better at dealing with outliers.

```
#Pre-processing 6 - Standardize the data

#Standardize the data - Min-max feature scaling (0, 1)
#scaler = MinMaxScaler(feature_range=(0, 1))
#LT = scaler.fit_transform(LT)

#Standardize data (0 mean, 1 stdev)
LT = StandardScaler().fit_transform(LT)
LTM = StandardScaler().fit_transform(LTM)
LTCM = StandardScaler().fit_transform(LTCM)
LCM = StandardScaler().fit_transform(LCM)
M = StandardScaler().fit_transform(M)

featuresSelected = []

featuresSelected.append(('Location, Time', LT, [], []))
featuresSelected.append(('Location, Time, Meteorology', LTM, [], []))
featuresSelected.append(('Location, Time, Canadian Indexes, Meteorology', LTCM, [], []))
featuresSelected.append(('Location, Canadian Indexes, Meteorology', LCM, [], []))
featuresSelected.append(('Meteorology', M, [], []))
```

Figure 17: Standardizing the data

6.a.4 Algorithms Compared

We decided to compare the models obtained with different classification algorithms (**Logistic Regression, Linear Discriminant Analysis, Nearest Neighbors, Naive Bayes, Classification and Regression Trees, Support Vector Machines**). We did this to try to find which one could better classify the existence of burned area.

We tried changing some parameters, such as the gamma for the Support Vector Machine, but ultimately only found some very small changes. As a result, we decided to stick with the default configurations for the most part.

```
#Prepare models
models = []

models.append(('LR', LogisticRegression(solver='lbfgs', max_iter=1000)))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(C=10.0, gamma=0.00001)))
```

Figure 18: Algorithms Compared

6.a.5 Software Used

For this phase of the project we decided to use Python as a programming language, for its easy to use machine learning capabilities. It is also a tool with appealing visualization options, as it is possible to create all sorts of different plots. Furthermore, it is something that we are very familiar with, which means we can benefit from our experience.

In order to utilize python, we resorted to Jupyter Notebook, since it has a very clean interface.

6.a.6 Results of the Study

6.a.6.1 Comparing Algorithms and Metrics

We started by using "accuracy" as the metric (0-100). In general, Classification and Regression Trees (CART) and K-Nearest Neighbors (KNN) had the best results, getting around 43 while the others tend to get 38 or lower. Though, it is worth noting that CART always got lower standard deviation.

	LT	LTM	LTCM	LCM	M
LR	37.368	37.368	37.368	37.368	37.368
LDA	37.564	37.564	37.564	37.564	37.564
KNN	**43.164**	**43.164**	**43.164**	**43.164**	**43.164**
CART	**42.952**	**42.756**	**42.568**	**42.756**	**41.399**
NB	38.733	38.733	38.733	38.733	38.733
SVM	26.022	26.022	26.022	26.022	26.022

Table 3: Comparison of the accuracy

We also tried using "balanced accuracy", $\frac{\text{sensitivity}+\text{specificity}}{2}$ **[7]**, which is good for unbalanced datasets such as ours (the number of fires with burned area is roughly 4 times less than the ones without).

With this metric the score gap between algorithms becomes much smaller, usually ranging from 40 to 44. However, the State Vector Machine still has by far the worst performance, getting exactly 35 on all setups. A look at its confusion

matrices allows us to deduce that it is always evaluating the existence of burned area as false, as it only contains True Negatives and False Negatives.

```
Confusion Matrix
[[274   0]
 [243   0]]
                precision    recall  f1-score   support

No Burned Area       0.53      1.00      0.69       274
   Burned Area       0.00      0.00      0.00       243
```

Figure 19: Example of a Confusion Matrix and some metrics for the State Vector Machine

We also tried using the f1 score, $2 * \frac{\text{Precision} * \text{Recall}}{Precision+Recall}$. In this case, Naive Bayes (NB) had the best results, with a slight margin over KNN and CART. The results were much lower for all the algorithms, which makes sense considering the metric used.

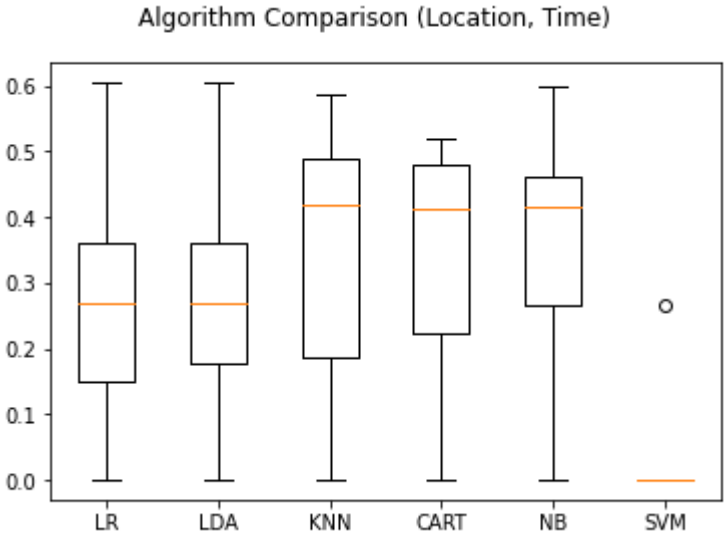

Figure 20: Box Plot for the selected features, using f1 score as the metric

In general, we can say that KNN, CART and NB were the best models for this scenario, as they had the best results. Meanwhile, the opposite can be said of the SVM, which had the worst results by far, regardless of metric chosen.

6.a.6.1 Comparing The Sets of attributes

We also compared the results across the sets of attributes. However, whether with accuracy or with the f1 score we only found differences for the "CART" algorithm between sets of attributes. LCM was the setup with the highest accuracy for that algorithm, but by a very small margin. So we can assume that this setup of location, Canadian indexes and meteorology is the best one for "CART".

6.b. Regression Study

6.b.1 Objective

Our objective is to create a model that can **predict** the **burned area** for **future fires**, as a way to assess the risks associated with certain conditions.

6.b.2 Source Data Model and Sets of Attributes used

The Source Data Model, Unused Attributes and Sets of Attributes used are the same ones from 6.a.2.

6.b.3 Data Preparation Activities

Note: A lot of what we did here was repeating the same steps from 6.a.3, as a lot of ideas are similar. That being said, there are some key differences.

Pre-processing 1 - Fixing the issue of having fewer rows in the dimensions – No differences.

Pre-processing 2 - Convert 'day_week' and 'month' so they are usable – No differences.

Pre-processing 3 - Inverse of ln(x+1) – On top of what was mentioned in 6.a.3, we also want to obtain comparable results to the study in **[5]** (which only did this to the burned area and later inversed it as well), to allow for a fair comparison.

Pre-processing 4 - Separate array into input and output components – No differences besides using the "burned area" as the output instead of the "boolean burned area".

Pre-processing 5 - Standardize the data – No differences.

6.b.4 Algorithms Compared

We compared these regression algorithms: **Linear Regression, Ridge Regression, K-Nearest Neighbors, Classification and Regression Trees** and **Support Vector Machines**. The last 2 were also used in the study in **[5]**, so they can serve us well in terms of comparison.

For these ones we chose to keep the default parameters.

6.b.5 Software Used

We once again resorted to using Python on Jupyter Notebook, for the reasons explained before in 6.a.5.

6.b.6 Results of the Study

6.b.6.1 Comparing Algorithms, Metrics and Sets of Attributes

Using the **positive Mean Absolute Error (MAE)** we realized that, regardless of the set of attributes chosen, the **SVM had the best performance** by a wide margin, always having an average of around 13, while the others all had around 20-30. Meanwhile, **CART had the worst results**, hovering around 30.

We can also see that the set of parameters M (Meteorology) had the best results by a fairly decent margin. This corresponds with the results obtained in the study this analysis is based on **[5]**, where the same was also verified.

In **comparison with the results from that study**, we had an **identical performance for the State Vector Machine**, which further confirms those

results. Meanwhile, "CART" showed very different results, having a much higher MAD in our case. This is probably due to a lack of fine tuning the algorithm parameters in our part.

	LT	LTM	LTCM	LCM	M
LR	19.555	20.578	21.044	20.744	20.201
Ridge	19.553	20.551	20.980	20.719	20.182
KNN	24.791	22.965	25.583	25.860	23.222
CART	27.172	31.300	29.392	27.245	25.217
SVM	**13.226**	**13.209**	**13.236**	**13.188**	**13.043**

Table 4: Comparison of the Mean Absolute Error (MAE)

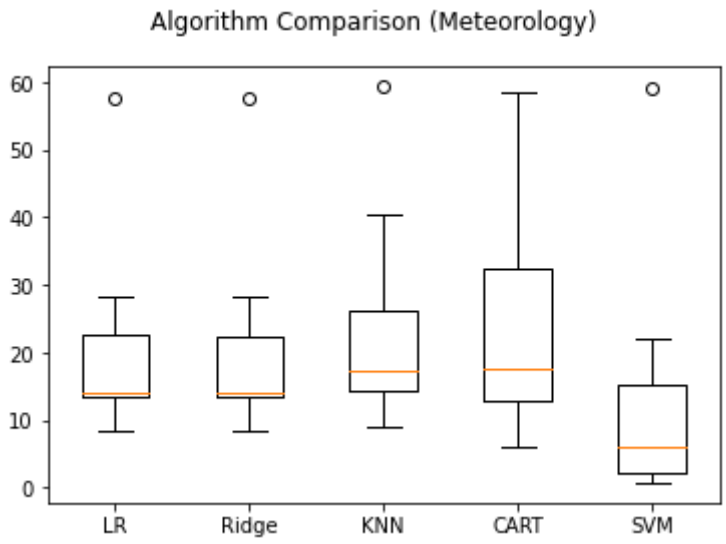

Figure 21: Box Plot for the Mean Absolute Error (MAE), with Meteorology as the set of attributes

Using the **positive Root Mean Squared Error (RMSE)** we obtained some similar conclusions to the previous metric. SVM still presents the best results and CART still has the worst ones.

As for the set of parameters, we can observe that M (Meteorology) is still the best, by a slight difference, with only an exception for the CART algorithm.

We can say that, in general, our results validate what was said in the study from **[5]**.

Regarding the comparison with the results from **[5]**, our CART model had relatively similar results (close to 64), even though it was not quite as good as the one they presented, especially for the set of attributes LTCM. Meanwhile, our State Vector Machine has seemingly better results than the ones from that study, though this would need further experiments to more thoroughly investigate if this is really the case.

	LT	LTM	LTCM	LCM	M
LR	42.659	43.441	43.287	42.820	42.426
Ridge	42.654	43.395	43.210	42.788	42.395
KNN	61.531	57.527	63.845	63.705	57.239
CART	68.042	67.922	78.816	67.710	70.563
SVM	**38.479**	**38.433**	**38.404**	**38.372**	**38.243**

Table 5: Comparison of the Root Mean Squared Error (RMSE)

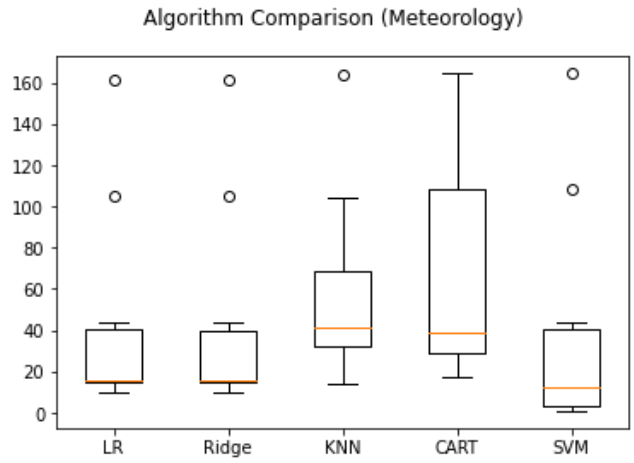

Figure 22: Box Plot for the Root Mean Squared Error (RMSE), with Meteorology as the set of attributes

References

1) «Forest Fire Area». https://www.kaggle.com/sumitm004/forest-fire-area (acedido Mai. 16, 2021).
2) P. Cortez e A. Morais, «A Data Mining Approach to Predict Forest Fires using Meteorological Data», p. 12. http://repositorium.sdum.uminho.pt/bitstream/1822/8039/1/fires.pdf (acedido Mai. 16, 2021).
3) N. R. Canada, «Canadian Wildland Fire Information System | Canadian Forest Fire Weather Index (FWI) System». https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi (acedido Mai. 16, 2021).
4) «The Best ETL Tools for Migrating to PostgreSQL», Severalnines, Abr. 18, 2018. https://severalnines.com/database-blog/best-etl-tools-migrating-postgresql (acedido Mai. 16, 2021).
5) «www3.dsi.uminho.pt/pcortez/forestfires/forestfires-names.txt». Acedido: Mai. 16, 2021. [Em linha]. Disponível em: http://www3.dsi.uminho.pt/pcortez/forestfires/forestfires-names.txt
6) «Log transformations: How to handle negative data values? - The DO Loop». https://blogs.sas.com/content/iml/2011/04/27/log-transformations-how-to-handle-negative-data-values.html (acedido Mai. 16, 2021).
7) «What is balanced accuracy? | Statistical Odds & Ends». https://statisticaloddsandends.wordpress.com/2020/01/23/what-is-balanced-accuracy/ (acedido Mai. 16, 2021).