

University College of Northern Denmark

Technology and Business

Computer Science Academy Profession (AP) Degree

Class: DMAJ0916 (Big Data)

Title: WASP Report– 4th Semester 2018

Project participants (Group 2):

-Ralfs Zangis

-Andrei-Eugen Birta

Supervisors:

-Nadeem Iftikhar

Due date: 2018-May-24 (14:00 UTC+1)

Submission date: 2018-December-17

Birta Zanigis

[1. Introduction 2](#_Toc513996170)

[2. Problem Statement 3](#_Toc513996171)

[a. Introduction 3](#_Toc513996172)

[b. Case Description 3](#_Toc513996173)

[c. Learning Goals 3](#_Toc513996174)

[3. Development Framework 4](#_Toc513996175)

[a. Pros 5](#_Toc513996176)

[b. Cons 5](#_Toc513996177)

[4. Development Process 5](#_Toc513996178)

[a. Data Acquisition 5](#_Toc513996179)

[b. Data Wrangling 6](#_Toc513996180)

[b.1. Identify and Handle Missing Values 6](#_Toc513996181)

[b.2. Data Formatting 8](#_Toc513996182)

[b.3. Data Normalization 8](#_Toc513996183)

[c. Descriptive Analysis 9](#_Toc513996184)

[d. Diagnostic Analysis 14](#_Toc513996185)

[e. Predictive Analysis 15](#_Toc513996186)

[e.1. Supervised Learning 15](#_Toc513996187)

[e.1.1. Chosen technique 15](#_Toc513996188)

[e.1.2. Technique implementation 15](#_Toc513996189)

[e.2. Unsupervised Learning 16](#_Toc513996190)

[e.2.1. Chosen Technique 16](#_Toc513996191)

[e.2.2. Technique Implementation 16](#_Toc513996192)

[f. Perspective Analysis 17](#_Toc513996193)

[5. Conclusion 17](#_Toc513996194)

[a. Denouement 17](#_Toc513996196)

[b. References 18](#_Toc513996197)

# Introduction

This document summarizes the collaboration of Group 2 for the Big Data specialization exam of the 4th Semester. The group consists of 2 members of 2 different nationalities. Because of the relatively small size of the group and because of the professional history of the group, the need of a working contract has significantly decreased and we managed to harness the benefits of diverse ideas and identify multiple possible approaches to certain problems, be it project or working regulations.

# Problem Statement

## Introduction

The purpose of this project is finding a way of reducing or down right preventing all collisions between, both civilian and military aircrafts, and wildlife animals, mainly birds, in the United States of America; by analyzing previous records of such events, and applying various Big Data analyzing techniques.

## Case Description

Day to day activities thought us that, the collision between something massive and something small and frail, usually ends up pretty bad for the small object and affects little to not at all the massive object, when it comes to aircraft collisions with birds and other wildlife creatures, things tend to go bad for both parties. Usually killing the animal and ruining the aircraft, possible for the rest of its “life”.

The following are images of possible damage that such a collision can cause, to an aircraft.

And considering that wildlife population is fluctuating depending on different seasons of the year, but usually increasing in numbers, such collisions should be taken with all seriousness and evaded as much as possible.

According to Allan and Orosz (2001) bird strikes cost commercial air carriers over US$1.2 billion worldwide from 1999–2000. Making the case, not only a safety issue, but also an economical one.

## Learning Goals

Some of the learning goals for this project are:

* Gathering useful datasets related to the case (Data acquisition)
* Converting datasets to a common format, in order to facilitate data analysis, using tools offered by Python
* Wrangling data (dealing with missing values, misspelled words or wrong datatypes), using tools offered by Python
* Describing what and why it has happened, using descriptive and diagnostic analysis techniques
* Predicting possible collision areas, at different times of the year and possibly wildlife species and types of damage, to facilitate warning emissions by concerned authorities

# Development Framework

Following previously acquired knowledge, from the 3rd Semester’s System Development course, we decided that the best way of choosing a development method is by evaluating the team and creating a Boehm and Turner Model.

The following image is the diagram we have come up with, following the self-evaluation process.



From the diagram above, resulted that we needed some kind of agile development method, due to the high amount of expected changes, small team size, and team’s culture, but is structured enough to accommodate for the project’s criticality.

Although the high level of criticality indicates that structured development method should be used, we have decided to work following the Kanban development method, because of two reasons: it best fits the other four measurements and it’s quite unlikely that our project will be used by any company, due to various reasons.

## Pros

* The small number of guidelines gives the programmer the freedom to work as he pleases
* The Kanban board helps manage the project responsibilities and keep track of what has been achieved
* The Kanban board allows for a better understanding of work and workflow.

## Cons

* Lack of “urgency” concept in the Kanban board, can lead to unnecessary waste of time, due to dependencies of certain tasks on other tasks

# Development Process

Typically, the development process for a Big Data project starts from one or more small and clearly defined questions, followed by Data Acquisition, Data Wrangling, Descriptive Analysis, Diagnostic Analysis, Predictive Analysis and ending with Perspective Analysis; all of them bringing important additions to the overall meaning of the project and helping those who are concerned about the matter, better understand the situation and take actions based on facts not on feelings.

## Data Acquisition

Data Acquisition is the first step that has to be made when working on a Big Data related project. This step refers to acquiring the necessary data for answering the previously defined question.

Our datasets were acquired from trusted websites that hold thousands and thousands of various datasets, the exact links for those datasets can be seen in the “References” part of this report.

The datasets we found and acquired are[[1]](#footnote-1):

* + USA Collisions from 1990-1999
  + USA Collisions from 2000-2009
  + USA Collisions from 2010-2018
  + USA Military Collisions from 1990-2018
  + USA flights in 2015
  + USA airports
  + USA airlines

## Data Wrangling

Second step in any Big Data related project, is Data-Wrangling or Data-Cleaning or Data-Cleansing. Although it is referred to under different names, they all denote the same actions, that being: cleaning and curing the data such that it will be ready to go for further processing.

We chose to do this, by using very powerful libraries, such as Pandas and Numpy, available for Python. Those libraries allow us to process the datasets in a much faster and reliable way than by doing it either manually or using other programming languages such as C# or Java.

The collision datasets were loaded into multiple data frames, using Pandas’s “read\_csv” function, then merged into one single data frame, just as the following image shows (fig. 1)



Figure

b.1. Identify and Handle Missing Values:

Missing values can mess with both the process and the conclusions that further processing of data will result in, thus handling them, is an important part that must be done as quickly as possible, in the development process.

But to be able to handle missing values, first we needed to identify them, then replace them with the standard NAN values. Following figure (fig.2) shows how we managed to achieve that.



Figure

We handled the, now filled in, missing values in different ways, depending on which technique would best fit the situation.

For example, we decided to go with a positive approach and fill N/A values in “nr\_injuries” and “nr\_fatalities” with 0, going with the presumption that if the incident would’ve had any casualties, someone, be it a reporter or staff member, would’ve looked and made sure that the data is recorded properly. Figure 3, will show exactly how we did that.



Figure

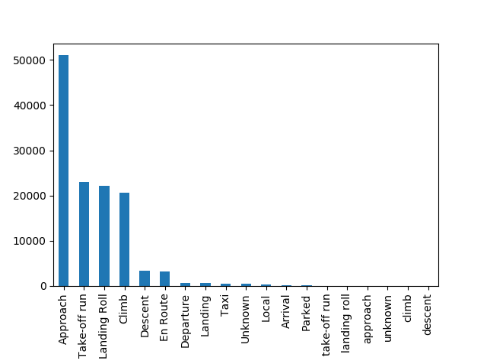
In some cases, we filled N/A values with the most frequent value, using “.mode”, because diagrams, like Figures 4, showed us that the data in those columns was mostly composed of same value, thus filling the missing values using “.mode” would result in a much smaller impact on the analysis.



Figure 4

Figure 5

In other cases, where we didn’t want the outliers to have a great impact on the analysis, we used “.median”(Fig.6).



Figure 6

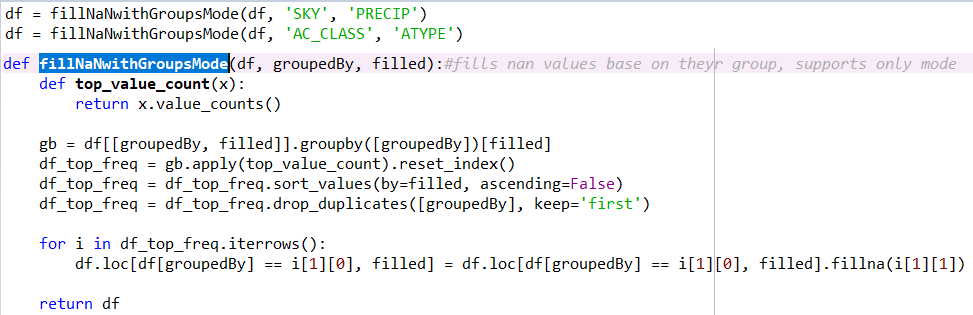
Columns, that store information like: precipitation, aircraft type or costs, required a much more complex approach.

Figure 7

Figure 7 shows how those “needy” column’s N/A values were dealt with, using another column’s information. First, the “helping” column was divided into several groups, then each group was examined and the value that appeared the most was used to fill in the unknown. Since every flight (row) had both “SKY” and “PRECIP” and since we now know the most common precipitation value for each SKY group, filling in the N/A did not represent as much of a problem anymore.

Or “COSTS”, which also required the creation of a special function, in order to fill it’s N/A values (Fig. 8). Both methods were created because we could not find any preexisting one that would yield the same results, in the Pandas library.

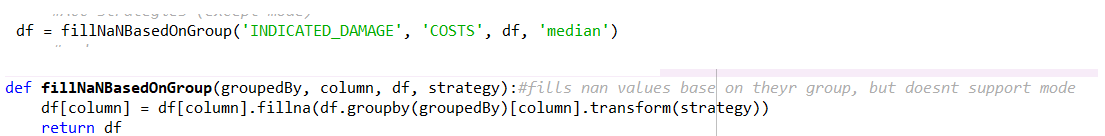


Figure 8

Some columns had to be dropped altogether, simply because were consisting of mostly N/A values and would bring little to no benefits in the next phases.

b.2. Data Formatting:

The purpose of Data-Formatting is to make sure that each column in the data frame, is of right type. This would help improve all the further processing done to the data frame, both from a performance point of view and a data-quality point of view.

Since we loaded the datasets into data frames, using special column types (Fig. 9, there was little to no additional work needed in this part.



Figure 9

b.3. Data Normalization:

Data-Normalization refers to making sure that all data is within the same range, to make sure that there are no misspells or wrong values in the wrong places.

One of the more important parts in this section, was making sure all the date stamps are saved in same format and in same field (as opposed to having 1 field for the date and one for the time). We achieved this using Pandas’s “to\_datetime” function. Another problem which we identified here was: sometimes the time was recorded with “dusk” or “dawn” or other words that can describe the time of day, instead of an actual hour. This was dealt with by replacing those values with the most common hour for each time of day, described by the specific word.

Another important part of this section, was making sure the airport names are not misspelled or wrongly typed in any way, shape of form, which we achieved by checking all the values within the column and “.replace” the wrong values with the appropriate ones.

-Indicator variables or dummy variables (columns that hold only LABEL info, not real data; ex: gas-type is a label, not real data)

## Descriptive Analysis

Descriptive analysis is the part in a big data project, where the Data scientist takes a look at the archived data and makes sense out of it. He analyzes it, creates charts and figures, that would ultimately lead him and all concerned parties, into achieving their ultimate goal. In our case, trying to reduce or down right stop future collisions between aircrafts and wildlife.

So, what can we see from previous events?

Certain things can be observed from the datasets we have acquired, the following picture(Fig.10) shows that, although we’ve seen a major decrease in recent times, this type of accidents follows a growing trend and every year more and more collisions happen.

Figure 10

We can also observe a seasonal pattern, Figure 11 is a combination of data from(1990-2016), divided per month, and shows that most accidents happen during mid-late summer and early-mid fall. And although these seasons have the most accidents, the number of injuries (represented in green), stays relatively the same. Where in case of fatalities(represented in black), is well below average.

Figure 9, shows us that most (more than 50%) of the collisions happened at 50 feet(~16 meters) or less, meaning that most collisions happened before or during take-off or landing. This fact only accentuates the importance of this case and how much of an impact a collision has, even though it happens at ground level.

Figure 11

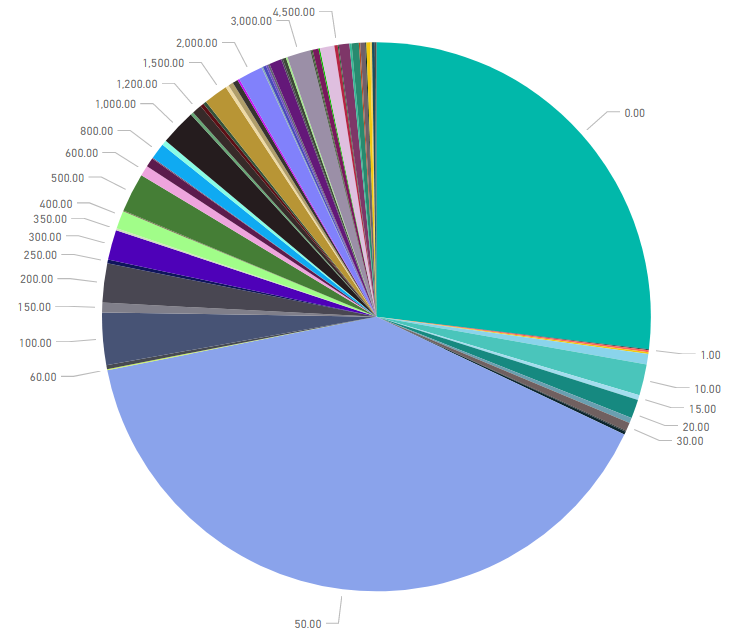
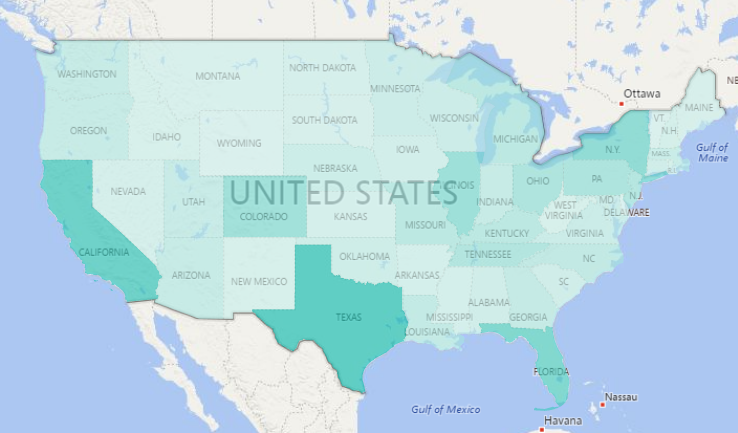
Airports tend to work according to the residing state’s laws, meaning they all follow the same set of rules of conduct. Figure 10, is a heat map, showing which state suffers the most from collisions with wildlife. Comparing to Figure 11 (which shows the busiest states, in terms of flights), we can see that, although … are the busiest states, … are also not the ones which have the most accidents.

Figure 10

Figure 12 (Accidents per aircraft altitude) (in feet)

Taking a deeper look at the previous finding, Figure 12 is a table which represents the busiest airports. While Figure 13 is a table which represents the airports with most collisions.

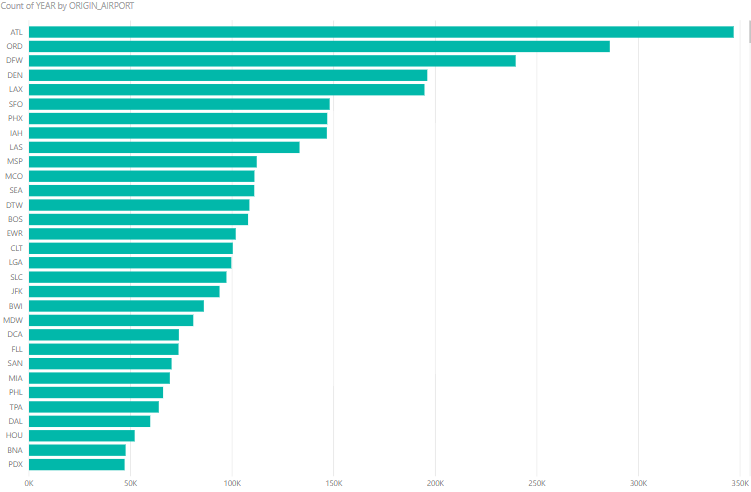


Figure 12 (flights per airport)

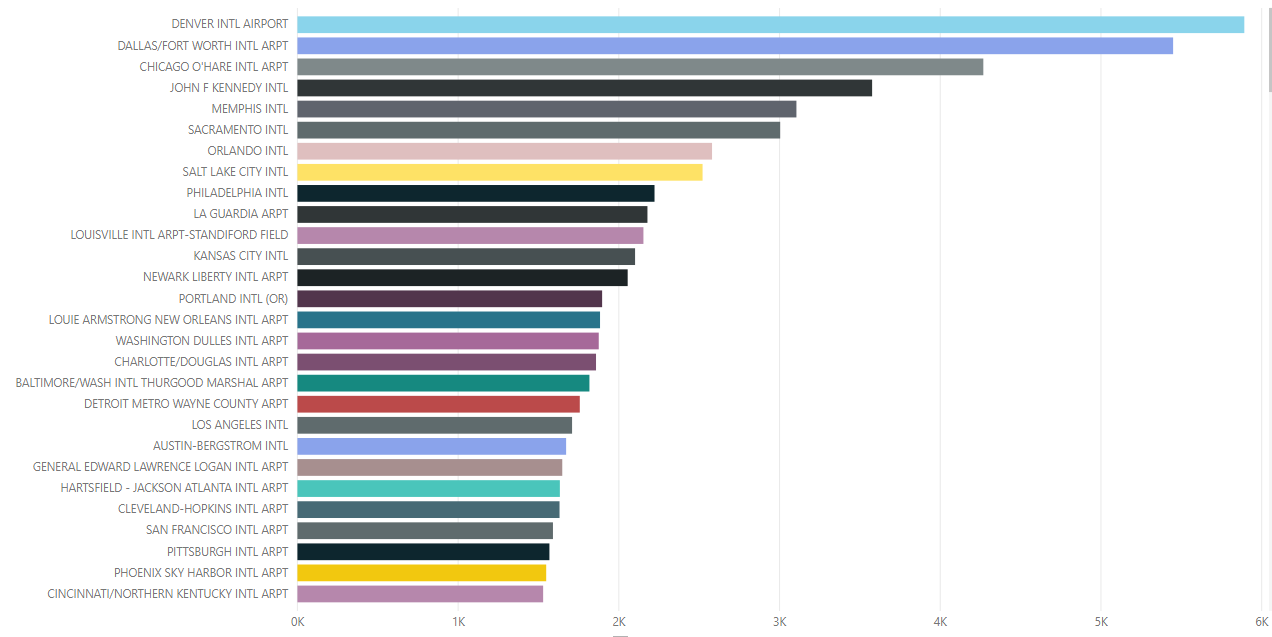


Figure 13 (collisions per airport)

We can see that although the DEN is the 4th busiest airport in USA, it has the most collisions with wildlife and ATL (Hartsfield–Jackson Atlanta International Airport), the busiest airport in USA, ranks place 23rd on the collisions list.

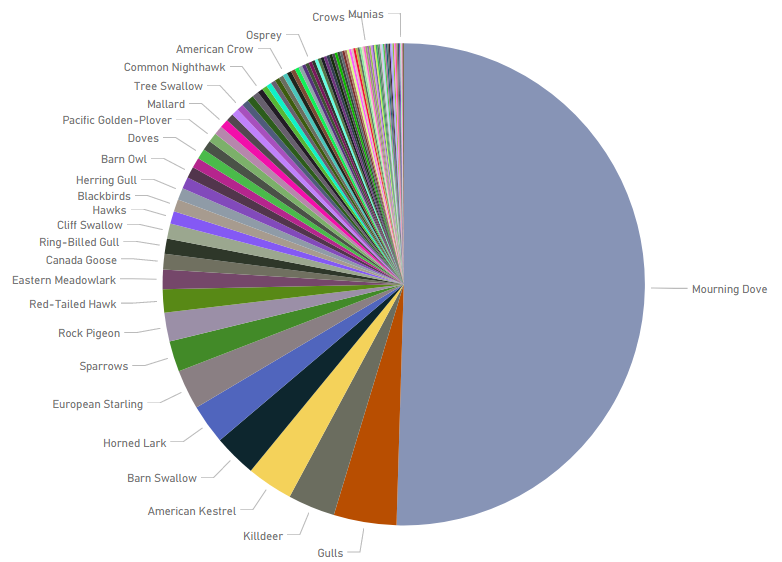
Figure 14 shows which are the most common species, aircrafts collide with.

Figure 14

Clearly showing that some species are more likely to partake in a collision, than others.

## Diagnostic Analysis

Diagnostic Analysis is the part in which the Data scientist finds the reasons why did the identified issues, in Descriptive Analysis, happened?

Why did those planes crash?

Well, the answer to that question is none other than wildlife. If there were to be no wildlife to collide with the aircrafts, no crashes would’ve happened.

* But how come, more than 50% of the collisions happen at altitudes bellow 16 meters?
* And how come some of the busiest airports and states are not the same as those that have the most crashes?
* Why are some species more likely to hit an aircraft than others?

In order to answer those questions, we needed more information.

According to a book written by Kaufman Kenn (“Lives of North American Birds”, 1996), Mourning Doves, the bird species which was part of 47% of all collisions since the 90’s, like to spend their time at altitudes between 5 feet (1.5meters) and 25 feet (7.6 meters), giving the reason why the Mourning Doves are more likely to crash into a plane, than any other species. “Adding salt to the wound” next species with the highest collision rate, are diurnal birds, natural predators to the Mourning Dove, be it for their meat, nests or eggs.

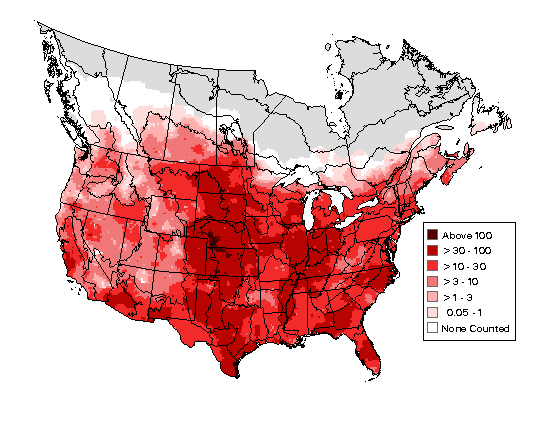
Figure 15 represents a heat map of Mourning Doves estimated population, over the USA territory. Comparing this one with Figure 10, we can see a clear resemblance between the two, showing part of the reason why some of the states have more collisions with Mourning Doves, than others. And although not all of the collisions are with Mourning Doves, next species that are likely to be part of a collision are natural predators to the Mourning Doves, and where there’re Doves, there’re predators as well, further increasing the risk that a plane will crash into a bird.

Figure 15 Breeding distribution and relative abundance of the Mourning Dove in North America based on the federal Breeding Bird Survey from 2011 to 2015 (Sauer et al. 2017).

## Predictive Analysis

Will it happen again?

Predictive analysis encompasses a variety of statistical techniques, but we chose to use machine learning to analyze current and historical facts, which would allow us to make predictions about future or otherwise unknown events.

In order to predict whether a flight will be part of a collision accident, we used a 2nd dataset which contained all flights in 2015, dataset which had ~6 million entries of flights on USA grounds. After cleaning the dataset and identifying which of those flights were crashes (using the initial dataset, which contained only crashed flights), we proceeded with the predictive analysis.

Our choice regarding this project was between supervised and unsupervised learning, since this project was partly made for learning purposes, we chose to implement both, to see what each of them is good at.

### e.1. Supervised Learning

Supervised learning maps an input to an output based on example input-output pairs. It uses labeled data consisting of features (X- input object) and labels (y- a desired output).

#### e.1.1. Chosen technique

There are multiple ways of doing Supervised Learning, and out of all of those, we decided to use decision tree, since it came out as best among the other techniques which we considered.

Naïve Bayes was one of the candidates for this part, but because it is biased towards common results, and since the chance of a plane partaking into a collision with wildlife is very small, we considered that the prediction accuracy would take a huge hit, were we to use this technique.

Logistic Regression was another candidate for this part, but because it is used to estimate discrete values based on given set of independent variables and because we wanted a concrete answer (whether the flight will crash into wildlife or not) as opposed to “There is 60% chance the flight will crash”, we decided not to implement this one.

#### e.1.2. Technique implementation

As shown in Figure 16, we started solving the problem, using decision tree, by importing necessary libraries for the job. Then separated the data frame into features and labels. After that, using “cross\_vaidation” we split the data into two parts (70% for training and 30% for testing). After this we trained the algorithm and then tested the results receiving the score of 99,53% accuracy.

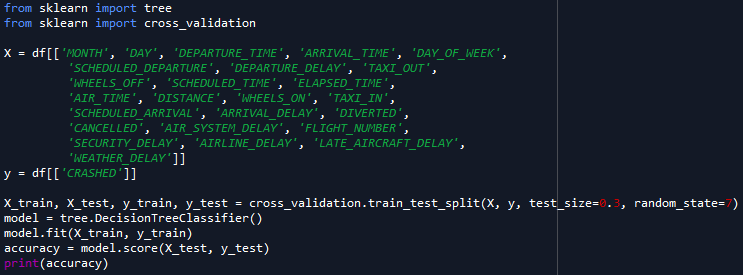


Figure 16

### e.2. Unsupervised Learning

Unsupervised learning describes hidden structure of "unlabeled" data. There are 3 main types of unsupervised learning techniques:

* Clustering: used to form groups in such a way that all individual entries in a group are similar to one another.
* Anomaly detection: used to identify anomalies in a system. Ex: bank frauds.
* Neural networks: system which is capable of learning. Ex: Image recognition, speech recognition, etc.

#### e.2.1. Chosen Technique

There are multiple ways of doing Unsupervised Learning, and out of all of those, we decided to use a clustering technique called K-means, since it came out as best among the other techniques which we considered.

#### e.2.2. Technique Implementation

As shown in Figure 17, we started implementing the K-Means, by importing the necessary libraries. Then using “cross\_validation” we split the data into two (70% for training, 30% for testing).

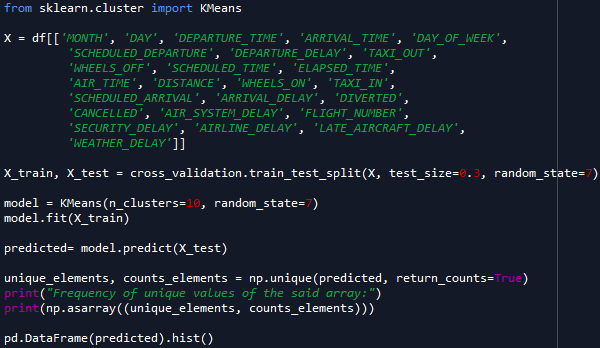


Figure 17

## Perspective Analysis

What can be done, in order to stop it from happening?

# Conclusion

## Denouement

In conclusion, during this semester we managed to not only gain knowledge about various Big Data techniques and a new programming language (python), but also a completely new part in the Programming world, a part which combines both programming and business into a concept which helps companies all around the world to learn about their customers, about what they are doing well and what not, and helping them making informed decisions instead of guessing over gut feelings.

Our project, turned out to be quite close to what we imagined when we were pitching the idea. We learned a lot of interesting facts regarding planes, birds and airports, while researching the matter. We also strengthened the knowledge gained during the courses and Digital Days event.

Although there is room for improvement, for example: the Predictive analysis could be further improved by using other datasets, like: bird flight patterns, weather patterns and a much bigger dataset of both successful and crashed flights; we are satisfied with what we achieved.

To see how we worked and what files we created, one has to follow the link, which will take you to our GitHub repository: <https://github.com/RaidenRabit/WASP>

As an ending note, we would like to thank all the readers, who invested their time in reading this paper, also the guiding teacher, who helped and guided us throughout the entire process. All files used in the creation of this report are attached to the hand-in folder, in case you would like to inspect them in great detail.

## References

* Diagram taken from “Balancing Agile with Discipline” by Barry Boehm Richard Turner
* https://wildlife.faa.gov/databaseSearch.aspx - the original dataset
* https://www.kaggle.com/faa/wildlife-strikes - dataset from kaggle (could be used as example)
* https://www.transportation.gov/ - for more info about transport
* https://wildlife.faa.gov/downloads/StrikeReport1990-2012.pdf - info about bird strikes (could be used in report and presentation)
* http://aircharterguide.com/Operators (filling blank info)
* Busiest airports: <https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/passenger/media/cy14-commercial-service-enplanements.pdf>
* Kaufman, Kenn (1996). Lives of North American Birds. Houghton Mifflin. p. 293. ISBN 0-395-77017-3.
* Mourning dove heat map: https://mnbirdatlas.org/species/mourning-dove/

1. each of the following is a different .csv file [↑](#footnote-ref-1)