

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](#_Toc514099935)

[2. Problem Statement 3](#_Toc514099936)

[a. Introduction 3](#_Toc514099937)

[b. Case Description 3](#_Toc514099938)

[c. Learning Goals 3](#_Toc514099939)

[3. Development Framework 4](#_Toc514099940)

[a. Pros 5](#_Toc514099941)

[b. Cons 5](#_Toc514099942)

[4. Development Process 5](#_Toc514099943)

[a. Data Acquisition 5](#_Toc514099944)

[b. Data Wrangling 6](#_Toc514099945)

[b.1. Identify and Handle Missing Values 6](#_Toc514099946)

[b.2. Data Formatting 8](#_Toc514099947)

[b.3. Data Normalization 9](#_Toc514099948)

[c. Descriptive Analysis 10](#_Toc514099949)

[d. Diagnostic Analysis 15](#_Toc514099950)

[e. Predictive Analysis 16](#_Toc514099951)

[e.1. Supervised Learning 17](#_Toc514099952)

[e.1.1. Chosen technique 17](#_Toc514099953)

[e.1.2. Technique implementation 17](#_Toc514099954)

[e.2. Unsupervised Learning 18](#_Toc514099955)

[e.2.1. Chosen Technique 18](#_Toc514099956)

[e.2.2. Technique Implementation 19](#_Toc514099957)

[f. Prescriptive Analysis 21](#_Toc514099958)

[5. Conclusion 22](#_Toc514099960)

[a. Denouement 22](#_Toc514099962)

[b. References 23](#_Toc514099963)

# Introduction

This document summarizes the collaboration of Group number 2, for the Big Data specialization exam of the 4th Semester. The group consists of two members of two different nationalities. Because of the relatively small size of the group and because of the professional history of the group members, working together; 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 related.

# 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 to them.

## 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. But 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 in the year between 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:

* Managing to gather 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 whether a future flight is going to partake in such an accident
* Prescribing proven ways of countering or reducing the likelihood of such events happening.

# 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 the likelihood of this project being used anything beyond examination reasons, is quite small.

## Pros

* The small number of guidelines gives the programmer the freedom to work as he pleases
* The Kanban board helps to 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 Prescriptive 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

Some datasets were acquired for direct purposes, like USA collision datasets, which have been acquired in order to facilitate Descriptive and Predictive analysis, where others, like USA airports and USA airlines, have been acquired in order to better understand the current data and to help in the Data Wrangling process.

## 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 in such ways 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 (Figure 1).



Figure 1

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 from among the others. Then replace them with the standard NAN values. Following figure (Figure 2) shows how we managed to achieve that, using Pandas’s “.replace” function.



Figure 2

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 3

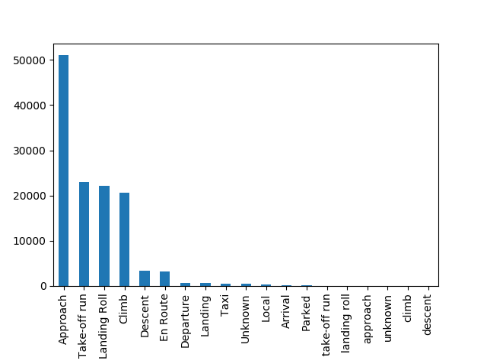
In some cases, we filled N/A values with the most frequent value, using “.mode”, because diagrams, like Figure 4, showed us that the data, in those columns was mostly composed of same value, thus filling the missing values using “.mode” (as can be seen in Figure 5) would result in a much smaller negative impact in future 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”(Figure 6). This decision allowed us to better fill in nan values, especially in places where there are a lot of “0” values, like in columns shown in image.

Figure 6

Columns, that store information like: precipitation, aircraft type or costs, required a much more complex approach, as they are more correlated with other column and to keep this correlation relatively the same when filling in nan values we decided to fill them by groups.

“COSTS” column was one of the ones effected in before mentioned way, as we noticed that depending on value in “INDICATED\_DAMAGE” the cost also changed, so we created simple method to fill in N/A values based on median in each group (Figure 7).

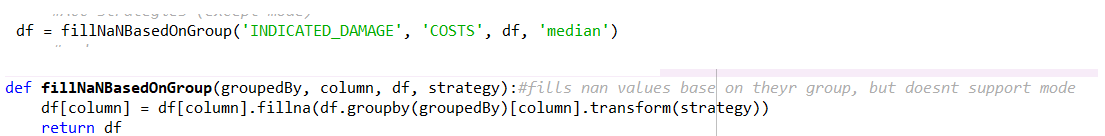


Figure 7

As figure 8 shows, previous method couldn’t satisfy all our requirements, as although the previous method could fill in N/A values based on group’s mean and median, it couldn’t do it using mode, so this this method was created to achieve it. 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.

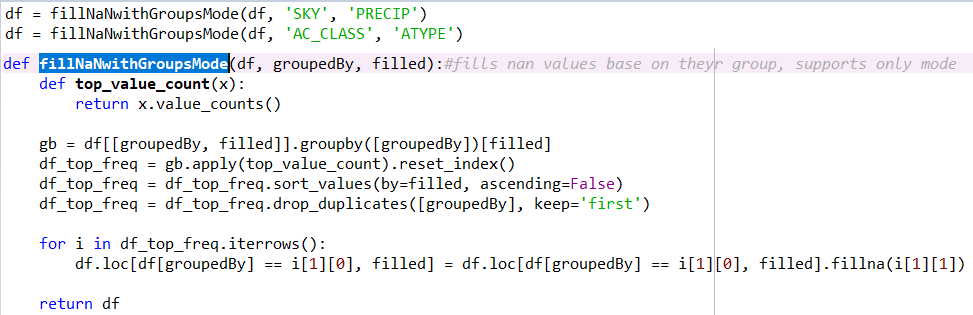
Some columns had to be dropped altogether, simply because were consisting of mostly N/A values and would bring little to no benefits in the next phases, while removing them would bring much higher performance related benefits. Or because they contained duplicate data, like “Incident\_month” and “Incident\_year”, data, which is already stored in “Incident\_date”.

Figure 8

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 (Figure 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-time 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 using the USA airport dataset.

b.4. Indicator variables or dummy variables

Indicator variables or dummy variables are columns which hold “label” information, for example: “gas-type” is a dummy variable, which tells us the type of fuel that an engine requires to run, where “km-per-liter” is a column that holds “feature” data.

In our case, we have several columns which serve as dummy variables, some of which are: “Type-Eng” (tells the type of engine), “Str\_Wing\_Rot” (struck wing), “Str\_Windshield” (struck windshield) and “Str\_Tail”(struck tail) and many more.

## Descriptive Analysis

Descriptive analysis is the part in a big data project, where the Data scientist takes a look at the archived, cleaned 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 final 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(Figure10) 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 12, 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

We’ve also found that around 90% of the collisions happen at or nearby airports.

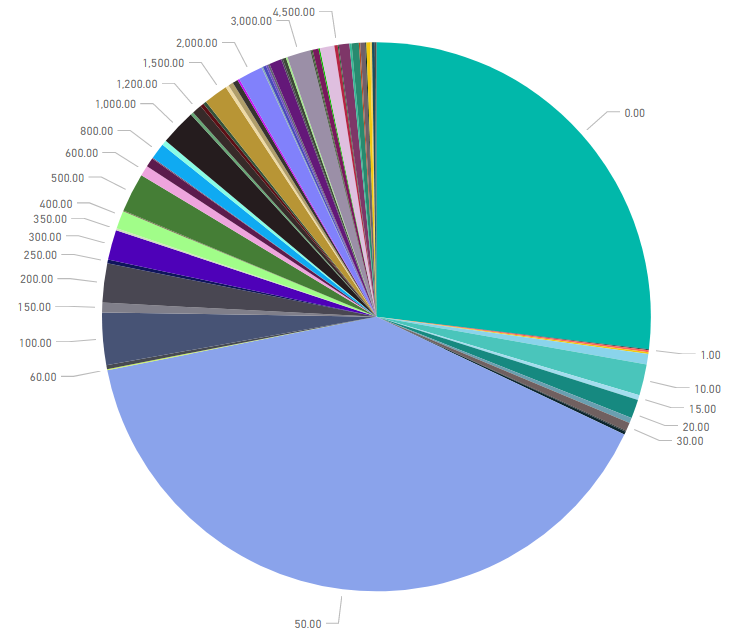
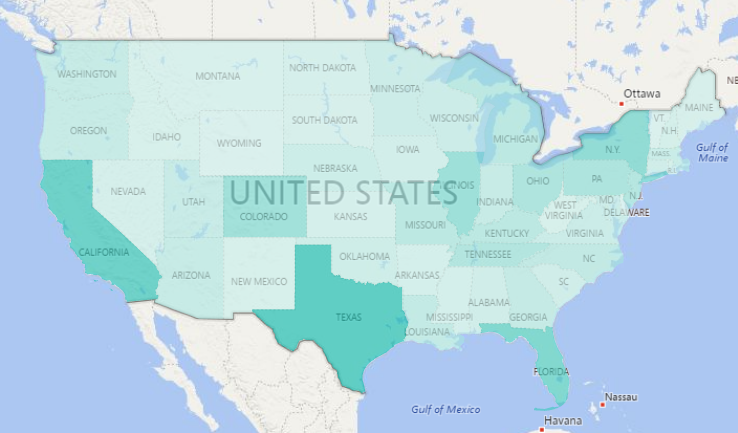


Figure 13

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

Figure 13, is a heat map, showing which state suffers the most from collisions with wildlife. Comparing to Figure 14 (which shows the busiest states, in terms of flights), we can see that while some states, like Texas and California are expected to have more collisions, due to their big number of flights, other states, like Nevada and Utah don’t seem to be affected by that at all.

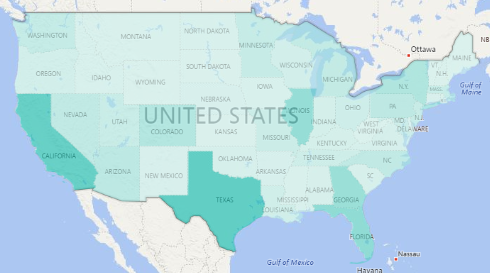


Figure 14 (busiest states by flights)

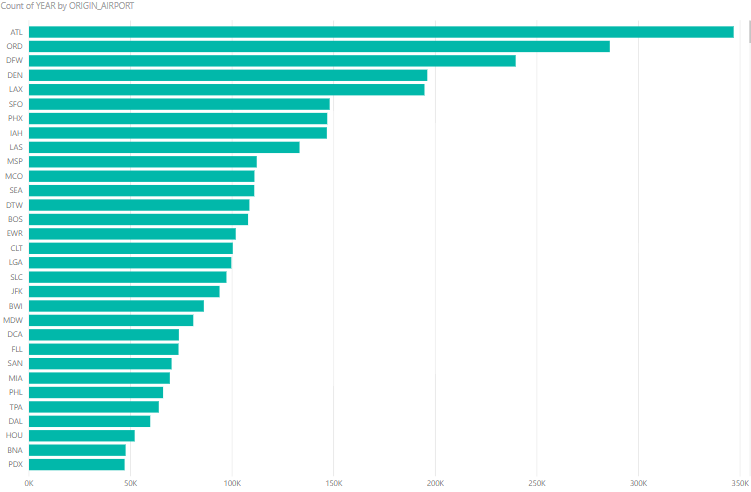
 Taking a deeper look at the previous finding, Figure 15 is a table which represents the busiest airports. While Figure 16 is a table which represents the airports with most collisions.

Figure 15 (flights per airport)

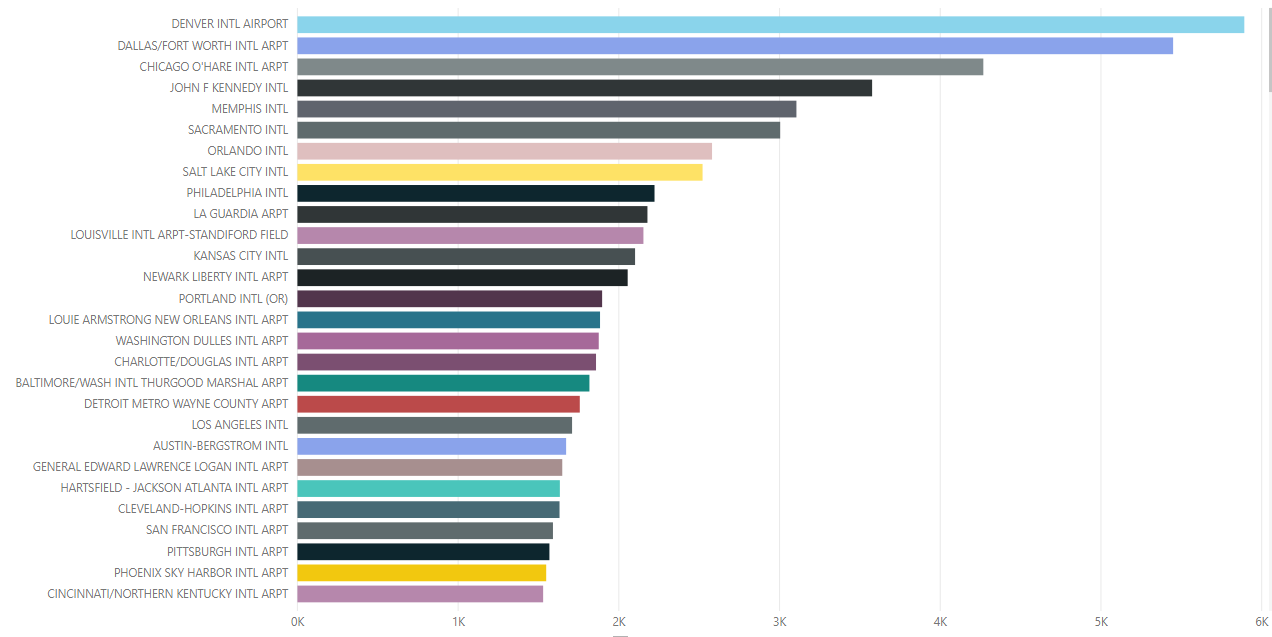


Figure 16 (collisions per airport)

We can see that although the DEN airport, 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, meaning that the number of flights, may not have much of an impact on the collision rate.

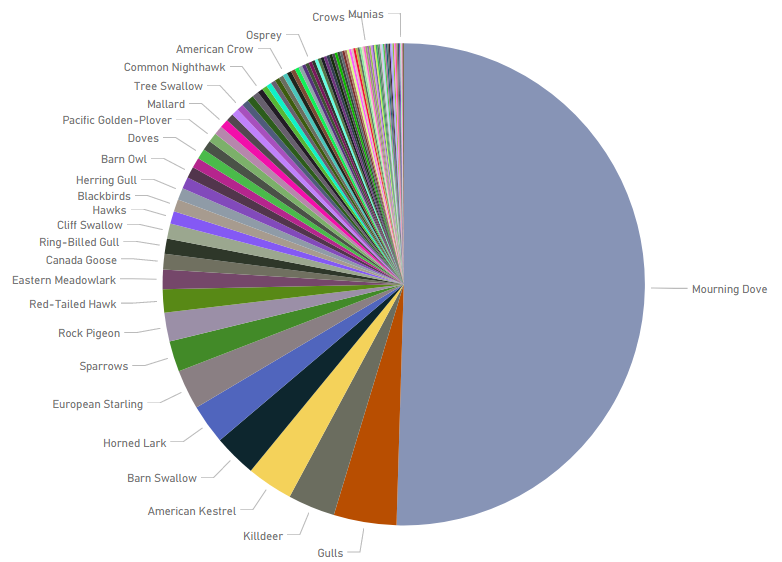
Figure 17 shows which are the most common species, aircrafts collide with.

Figure 17

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

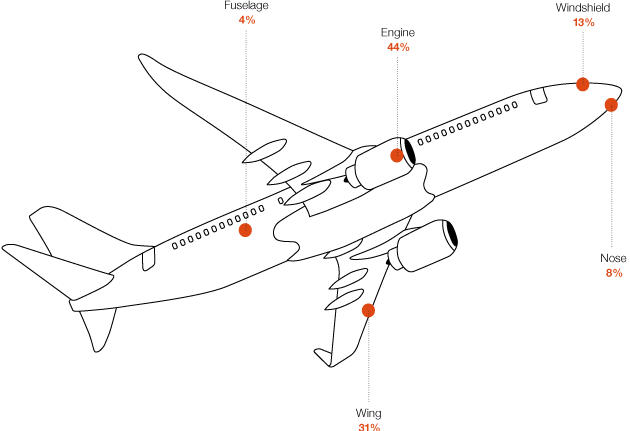
Figure 18 shows the distribution of bird strikes onto a typical aircraft. 44% of the collisions affect the engine rendering the aircraft useless for a period of time tightly dependent on the aircraft type and the availability of replacement parts. The wings and the engine are the only parts of an aircraft that hold fuel, and a collision to either of them, can cause the engines to explode, putting in danger the passenger’s lives or causing fuel leakages, forcing to aircraft to execute an emergency landing.

Figure 18

## Diagnostic Analysis

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

* How come, more than 50% of the collisions happen at altitudes bellow 16 meters?
* 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.

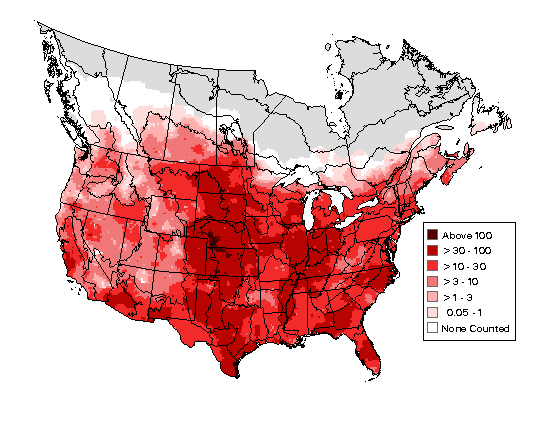
Figure 19 represents a heat map of Mourning Doves estimated population, over the USA territory. Comparing this one with Figure 13, we can see a clear resemblance between the two, showing part of the reason why some of the states have more collisions with Mourning Doves, than others. And although not 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 19 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

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.

Will it happen again?

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). Using this method, we managed to create a model that is capable of predicting, whether a flight will partake in a wildlife collision or not, with a quite astonishing accuracy.

* Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
* Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

#### 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.

Other techniques which we considered are:

* 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 20, we started by separating data frame into features and labels. Than we decided how many rows of this dataset we would like to predict, then followed by using cross\_vaidation to split this for training purpose (70% training/30% testing). After this we trained the algorithm and then tested the results receiving the score of 99,53% accuracy. Although the resulting score may seem very accurate, we determined that it is this precise only because the possibility of animal collision is tiny, making the algorithm very biased towards plain not crashing. In the end we decided to predict few rows of dataset of original dataset, which wasn’t used for training or testing, to base our predictions on and in the end compared the received results to expected ones.

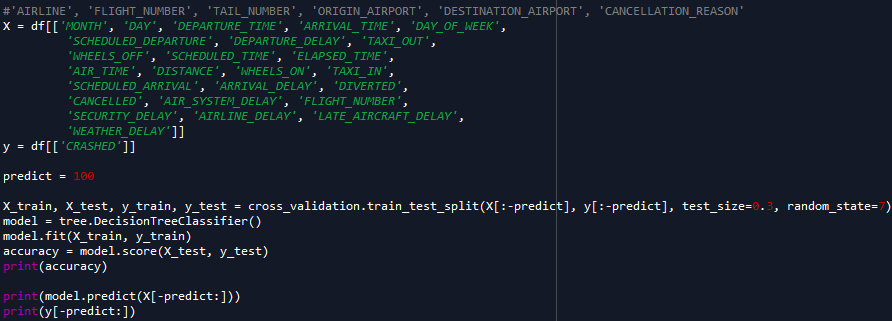


Figure 20

### e.2. Unsupervised Learning

Unsupervised learning describes hidden structure of "unlabeled" data. Since the examples given to the learner are unlabeled, there is no evaluation of the accuracy of the structure that is output by the algorithm, it is used to predict pattern in data —which is one way of distinguishing from supervised. 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. The reason we chose to do so, is not only because K-means is one of the, if not the most popular technique, used in situations similar to this one, but also because K-means would help us in improving the Supervised learning model, previously mentioned, by grouping the flights into several clusters, further curing the data and forcing the model to predict based on available data, for similar flights, as opposed to predicting based on all data, indifferent on how related, and possibly, useful it is.

Other techniques that we considered are:

* Hierarchical clustering, which is a method of clustering that seeks to group data using, either an Agglomerative(bottom-up) or Divisive(top-down) approach. Technique which we did not choose because, as many results on the internet showed, is slow and requires a lot of memory.
* Deep belief network, which is a type of neuronal network, that is composed of several layers connected with one another, which is capable of adapting and learning to better predict future outcomes. Technique which we did not choose because it required a deeper level of knowledge and understanding on the matter of neuronal networks, than we currently have.

#### e.2.2. Technique Implementation

We started the work on unsupervised learning, by (as seen in Figure 21) deciding the amount of values, the model, would be tested with. Then we decomposed the data into two components and proceeded to decide number of clusters (done using elbow analysis) and then thought the algorithm using the extracted information. Than we added the “clusters” column to data frame, so we can store the results of clustering. At the end we continued by predicting in which of the four clusters, the new values would fit.

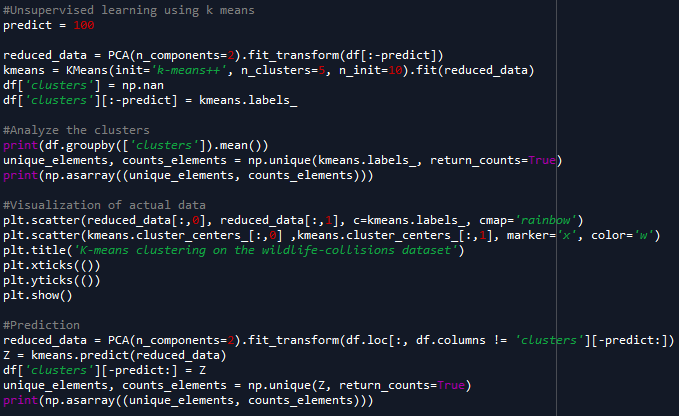


Figure 21

As the result of this codes execution visualized information about the clusters was displayed in easily understandable way (Figure 22).

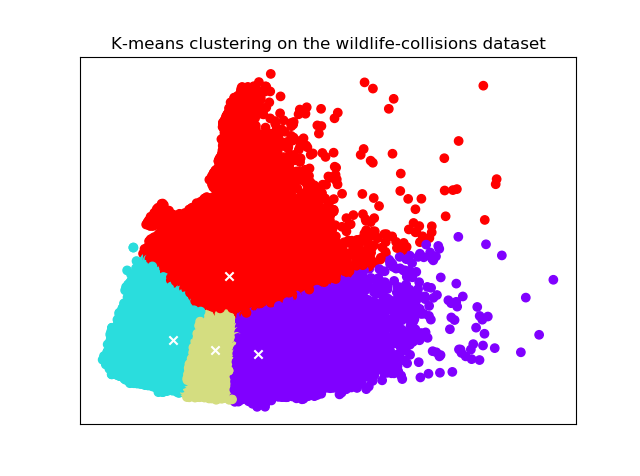
\

Figure 22

How did you decide to have four clusters?

To decide the number of clusters we had the choice of using either Silhouette Analysis or Elbow Analysis, we chose latter, as it was both simple to implement and work with. Chosen method looks at the percentage of variance with different number of clusters.

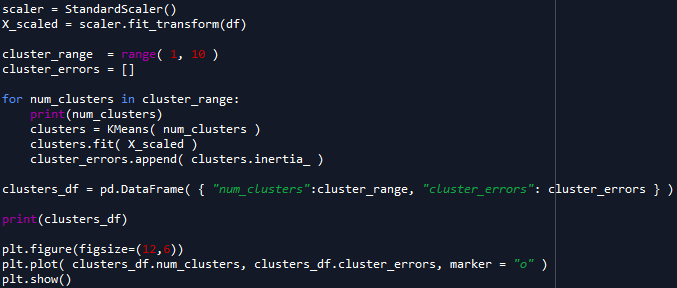
Figure 23, shows how we started work by scaling our dataset, so our data isn’t affected by outliners, making mean value 0 and standard deviation of 1. Following this step, we executed the algorithm 10 times with different number of clusters and in the end the percentage of variance was plotted (Figure 24) to visualize the results.

Figure 23

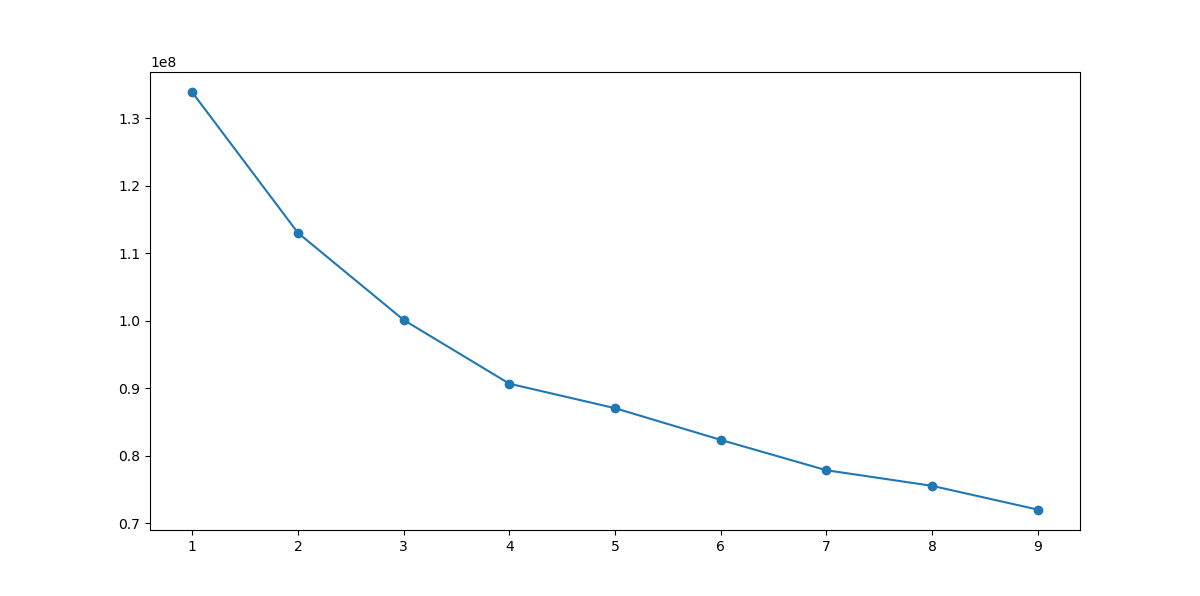


Figure 24, shows us that the variance, starts dropping slower at 4th cluster, which lead to our decision of using only 4 clusters for further processing.

Figure 24

## Prescriptive Analysis

The Prescriptive analysis is the analysis which gives the concerned authorities, a “prescription”, some advice on how to reach the goal or, in this case, how to prevent collisions from happening.

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

According to an article by Melissa Mayntz “Airport Bird Control Methods”, there are three main of minimizing airplane-wildlife collisions, most successful airports have reported using a combination of all three:

* Modifying the Habitat:
  + Removing seed-bearing plants to eliminate food sources
  + Covering nearby water sources with netting to prevent birds from landing
  + Removing brush and trees that serve as attractive nesting sites
  + Keeping grass mowed short so it is not as suitable for bird shelter
* Controlling bird behavior:
  + Using sonic cannons, recorded predator calls and other noise generators to disrupt birds
  + Using lasers at dawn and dusk to simulate predators and scare birds away
  + Flying trained falcons over roosting areas to disrupt birds before they nest
  + Training dogs to track through the habitat and teach birds that the area has many predators
* Controlling aircraft behavior:
  + Training spotters with binoculars and scopes to pinpoint hazardous bird’s areas and directing planes to different runways or approaches
  + Using radar equipment to track the movement and density of bird flocks to predict their behavior and manage control techniques more effectively
  + Adjusting flight times to avoid the busiest hours for bird activity, such as early morning and late evening or during peak migration periods

According to Roger Nicholson, Ph.D., Associate Technical Fellow, Aviation System Safety, and William S. Reed, Safety Pilot, Boeing Flight Technical and Safety; some of the best preventive measures against wildlife collisions are:

* Delaying the takeoff or landing when fuel permits.
* Take off or landing on another runway.
* Thorough preparation and training effectuated by all crew members, be it pilot or flight assistant.

According to Michael Begier, national coordinator of the airport wildlife hazards program at the US Department of Agriculture; the use of pyrotechnics, like flash bangs, gun shots or fireworks, has been proved to be quite efficient in distracting birds and scaring them away from an runway.

Ellen Lindblad director of planning and environmental compliance at Southwest Florida International Airport has employed Sky, a canine helper that managed to reduce the wildlife collisions at RSW airport by 17%, by simply chasing them away.

Salt Lake City airport (SLC) has employed the use of pigs in order to disrupt the habitat of several Californian Gulls families in order to scare the birds away from the adjacent to the airport island, that severed as the birds habitat.

# 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> - Dataset for collisions
* <http://aircharterguide.com/Operators> – Used to normalize data
* <https://www.kaggle.com/freddejn/flights/data> - Dataset for all flights
* <https://openflights.org/data.html> - info about airports and flight operators
* <http://www.awesomestats.in/python-cluster-validation/> - how to find Nr. of clusters
* <https://data.humdata.org/dataset/ourairports> - used to know state airport is in
* <https://wildlife.faa.gov/downloads/StrikeReport1990-2012.pdf> - info about bird strikes (could be used in report and presentation)
* 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/>
* Collision distribution over aircraft: <http://www.boeing.com/commercial/aeromagazine/articles/2011_q3/4/>

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