

A thick dark blue vertical bar runs down the left side of the page. A blue arrow-shaped banner points to the right from this bar, containing the date '4/12/2019'. In the bottom left corner, there are several thin, curved, light blue lines that sweep upwards and to the right.

4/12/2019

FAA Investigation into Wildlife Collisions 2000- 2012

Business Intelligence Visualisation

Eoin Dalton
20070289

Table of Contents

1	Introduction	2
2	Target Organisation and user	3
3	User role.....	3
4	User KPI's defined and explained.....	3
5	Data cleaning – actions taken	4
6	Understanding data	6
7	Data analysis - preparation and summarisation	8
8	Software.....	10
9	Limitations.....	12
10	Conclusion	13
11	References	14

Table of Figures

Figure 5.1: Formatting data values.	4
Figure 5.2: Formatting time value.....	5
Figure 5.3: Formatting blank cells.....	5
Figure 6.1: Pivot table with top ten airport collisions.....	6
Figure 6.2: Pivot table with top ten Airline carrier collisions.....	6
Figure 6.3: Total collisions by year.....	7
Figure 6.4: Part of the day most collisions happen.....	7
Figure 6.5: Time of day most collisions took place.	7
Figure 6.6: Number of collisions per species.	8
Figure 6.7: Number of collisions per aircraft.	8
Figure 7.1: Top 30 busiest Airports in America.....	9
Figure 7.2: Correlation between time and number of collisions.	9
Figure 7.3: Coefficient of variation	10
Figure 8.1: Loading XLS script.	10
Figure 8.2: Master calendar.	11
Figure 8.3: Aerospace dashboard.	12
Figure 8.4: Species dashboard.	12

1 Introduction

As part of our Business Intelligence Visualisation module we were giving an assignment in which we were asked to create a dashboard using QlikView which is a business intelligence visualisation tool. We were given a big dataset containing data about wildlife collision by aircrafts. There was information like times and date, location airports, aircraft operators, weather, damage, species and more. The dataset was given to us in XLS format and was fairly raw. What is meant by raw is some preparation was needed before the data was ready for QlikView.

This report will identify an organisation that may be able to gain insight from the data and will identify a user in which the dashboard will be built for. The objective will be to identify some Key Performance Indicators (KPI's) that the organisation will need to gain that insight. After the organisation, user and KPI's are identified the report will then take the reader through the steps involved in the data preparation/cleaning. Before QlikView comes into the equation, Excel will be used to understand the data. This will be done through Excel functions and pivot tables to summarise and analyse the data.

Once the data is prepared it will then be loaded into QlikView and a dashboard will be built that will focus around the KPI's that were identified. This part will show some of the logic behind the charts that were used. Finally, the report will finish with a short conclusion.

2 Target Organisation and user

A few different thoughts came to mind when investigating the dataset and trying to identify an organisation and user for this project. The first scenario was from the perspective of an airport operator. While gaining an understanding of the data through pivot charts there was some airports that had high numbers of collisions Denver was one that springs to mind. But the problem was the dataset got really small when focusing on one airport.

What was noticed was the high number of American airports on the list. After a Google search the top 30 busiest airports in America was identified and were all contained within the dataset. The Federal Aviation Administration (FAA) are the governing body that oversees all the aviation transport in America. The FAA is structured with a Deputy Administrator. Under the deputy administrator is five associate administrators also reporting to the deputy administrator is chief council and nine assistant administrators (FAA, 2019).

It was decided an investigation would be made into the number of collisions happening in America with the target organisation being the FAA. The user would be Deputy Administrator.

3 User role

After receiving a high number of complaints from the Wildlife Conservation Society about the high number of wildlife collisions by aircrafts in America the FAA needed to take some action. The deputy administrator who is the head of the FAA ordered an investigation into the complaints. The deputy administrator wanted to get a summarised view of the which airport and carriers was responsible for the highest number of collisions. The time and part of day most collisions were happening. What was the species that was hit the most? The job of obtaining this information was giving to two of the nine assistant administrators. They were tasked with building a dashboard with QlikView that would enable the deputy administrator to make decisions on if there was a way to reduce the number of collisions per year.

4 User KPI's defined and explained

As stated in the user role the deputy administrator was after summarised data. They stated what data they were after. The data they were most interested in was the number of collisions this was the basis of all other indicators. The following list is the main KPI's identified to build the dashboard.

- Number of Collisions
- Airport Identification
- Airline Carrier Identification
- Species Identification
- Part of the day
- Time of Day

Now let's look at the logic behind some of the KPI's that were identified. Airport Identification this is important to see which airport was having the most collisions. They may be a correlation between location of the airport and the type of species being hit. Airline

carrier was the next KPI. The deputy administrator needed this information so they could go back to the airline with evidence that could persuade them to maybe schedule flights at a different part or time of the day. Species identification was very important because if there is a high number of a certain species being hit the maybe they can be relocated to a suitable habitat this can reduce the amount of collisions happening. The part and time of the day needed to be identified because if there was a time of the day the most collision were happening maybe there is some correlation with the activities of a certain species then change may be made to flight times. All these KPI's are focused around the amount of collisions so this is the most important KPI's of all.

5 Data cleaning – actions taken

As part of the cleaning process there were a few problems identified such as date values and blank fields the format of time values was another problem. The first step in cleaning the data involved formatting date values. Seen in figure 5.2 the reported date field highlighted by the red arrow shows a date and time, these are formatted as general text highlighted by the red arrow at the top. By changing this value to a short date as seen further down the image the date value is now represented in a suitable way for QlikView to interpret by using DD/MM/YYYY.

The figure consists of two screenshots of a QlikView spreadsheet, illustrating the process of formatting date values. Both screenshots show a table with columns: W (Reported: Date), X (Wildlife: Siz), Y (Conditions: Sk), Z (Wildlife: Species), AA (When: Time (HHMM)), and AB (When: Time).

Top Screenshot: The 'Reported: Date' field in row 12 contains the value '02/01/2000 00:00'. A red arrow points from the 'General' format option in the dropdown menu to the 'Reported: Date' field.

Bottom Screenshot: The 'Reported: Date' field in row 12 contains the value '02/01/2000'. A red arrow points from the 'Date' format option in the dropdown menu to the 'Reported: Date' field.

W	X	Y	Z	AA	AB
Reported: Date	Wildlife: Siz	Conditions: Sk	Wildlife: Species	When: Time (HHMM)	When: Time
	Medium		Unknown bird - medium		
	Medium		Unknown bird - medium		
	Medium		Northern harrier		
02/01/2000 00:00	Small	Some Cloud	Unknown bird - small	1345 Day	
06/02/2000 00:00	Large	No Cloud	Canada goose		Night
	Small	No Cloud	Unknown bird - small	1530 Day	
	Small		Snow bunting	1539 Day	
	Small	No Cloud	Unknown bird - small	1722 Dusk	
	Small	No Cloud	Snow bunting	1220 Day	
	Small	No Cloud	Snow bunting	842 Day	
	Medium	No Cloud	Unknown bird - medium	1721 Night	
08/01/2000 00:00	Medium	Overcast	Unknown bird - medium	1450 Day	

Figure 5.1: Formatting data values.

The next field that need to be formatted was the time column. Referred to with the arrow on the left-hand side of the image is the time before the excel function was applied to it. This shows a number 1345 which should be 13:45 which is a much easier way to interpret. To format the date an excel function was used seen highlighted by the arrow at the top of the image (YouTube, 2016). The function works by converting text and formatting the cell Z9 as referred to in figure 5.2 then it shows the leading zero in between double quotes followed by another zero then because when you are doing custom number formatting, we can use the \ to insert the colon the two more zeros and close the double quotes. Because this is text, we want to convert it back to a number any maths function will do this, so by adding a zero it will convert this to a number. And as seen by the red arrow to the right all the times are now formatted in the correct way.

X	Y	Z	AA
Conditions: Sky	Wildlife: Species	When: Time (HHMM)	When: Time (HHMM)2
overcast	Unknown bird - medium	1345	09:43:00
	Unknown bird - medium		00:00:00
	Unknown bird - medium		00:00:00
	Unknown bird - medium		00:00:00
	Northern harrier		00:00:00
	Unknown bird - large	1300	13:00:00
Cloud	Unknown bird - small		00:00:00
overcast	Unknown bird - small		00:00:00
Cloud	Unknown bird - medium	1105	11:05:00

Figure 5.2: Formatting time value.

Another problem was the large number of blank cells as stated above. To format the blank cell excel provides a setting to find them. This process was a little pain staking as each column contained a number of blank cells and some values were text and somewhere numbers. This meant that each column needed to be formatted individually. Referred to in figure 5.3 is the setting to find the blank cells. To the right of the image is the column shown not data, after finding the blank cell excel allows you to replace or delete the blanks. By deleting the cell, we were losing to much valuable data so in this case it was replaced with No data it other cases it was replaced with zeros or text that was applicable to the situation.

G	H
Effect: Other	Location: Nearby if en route
No Data	No Data
No Data	No Data
No Data	No Data
No Data	No Data
No Data	No Data
No Data	No Data
No Data	No Data
No Data	No Data
No Data	No Data
No Data	No Data
No Data	No Data

Figure 5.3: Formatting blank cells.

6 Understanding data

In order to gain a good understanding of the data there was a series of pivot tables and graphs used. Looking at the data in a spreadsheet make it really difficult to understand what is happening. Referring back to the KPI's all the data that the chief administrator was looking for was surrounding the collisions that occurred in the given time-period. The first pivot table was to get an idea of the total collisions for the top ten airports as referred to in figure 6.1. You will see that between 2000-2012 Portland Intl had the highest amount of collisions.

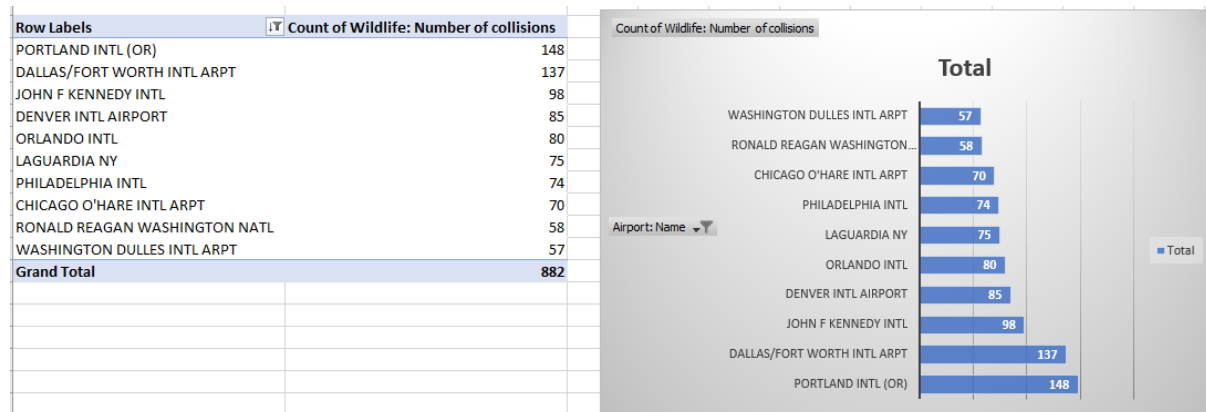


Figure 6.1: Pivot table with top ten airport collisions.

The next pivot table was enabling us to understand which airline carrier had the most amount of collisions. This pivot table was filtered down to the top ten with the most collisions. If you refer to figure 6.2 you will notice the American Airlines had the highest number of collisions followed by United Airlines.

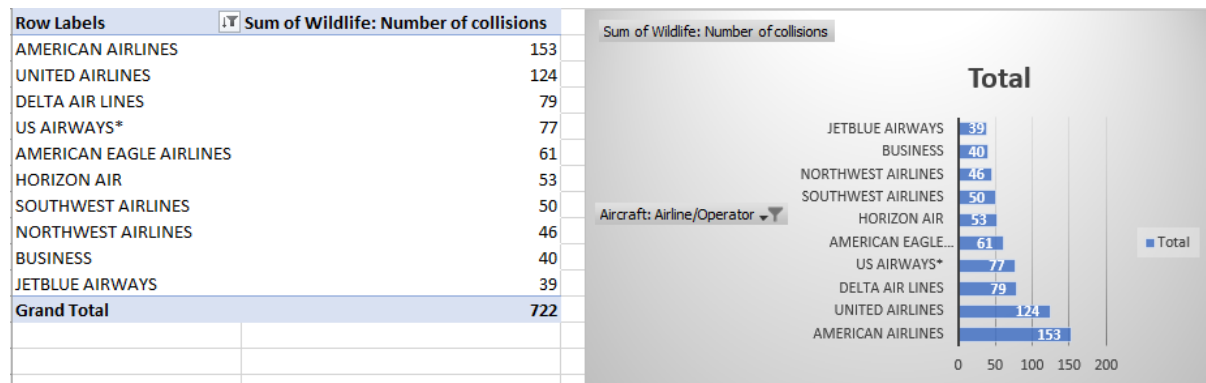


Figure 6.2: Pivot table with top ten Airline carrier collisions.

The data supplied was in a time period from 2000-2012 referred to in figure 6.3. By summing up the collisions and using the report date gives us an idea on the number of collisions that occurred for each year. As shown in figure 6.3 below 2002 had the highest amount of collisions.

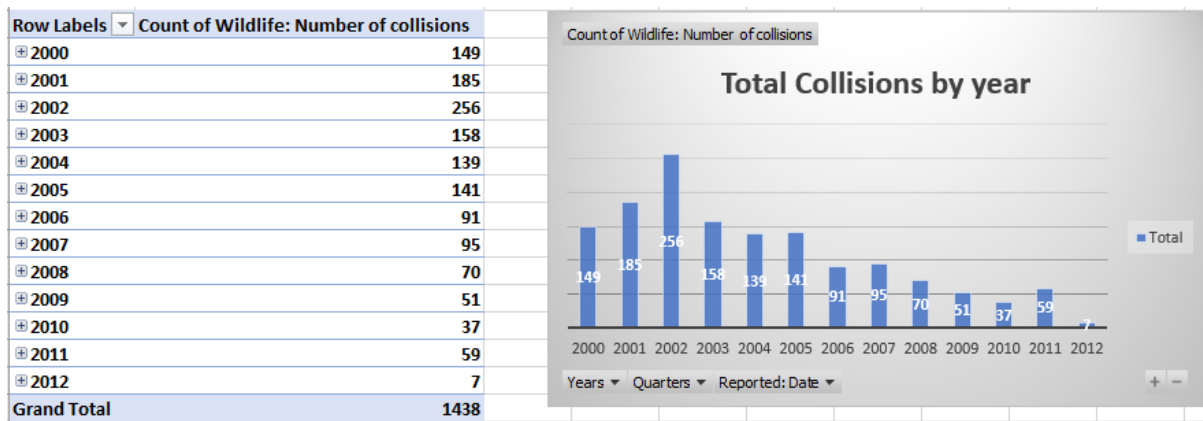


Figure 6.3: Total collisions by year.

In order to get an idea of what part of the day most collisions were happening there was two dimensions needed. The first is number of collisions, and the next is part of the day field. By doing this it was identified the day time had 914 collision as shown in figure 6.4. This significantly more than any other part of the day. This maybe something that could be looked at by the deputy administrator.

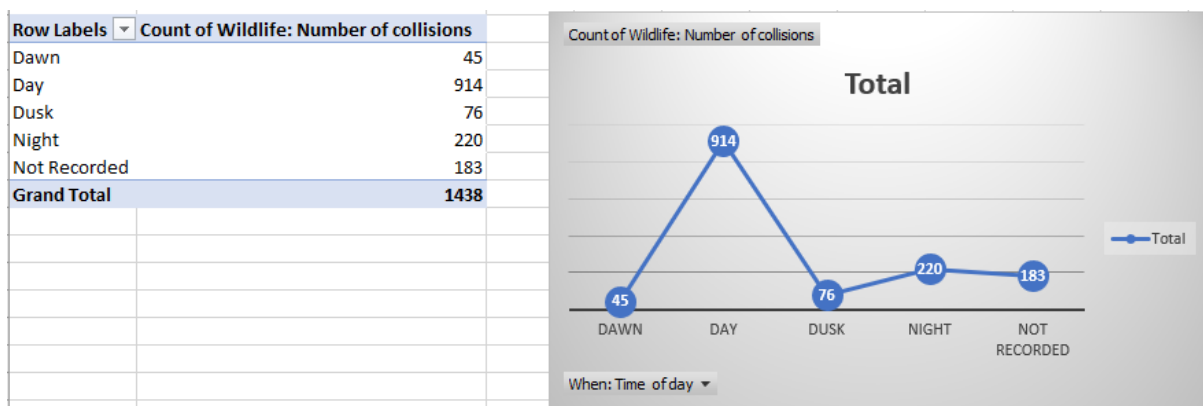


Figure 6.4: Part of the day most collisions happen.

In the figure 6.5 in the time of the day that most collisions were happening. It seemed that in the middle of the day there was 11 species hit. But looking at this other figure there doesn't seem to be much correlation between time and the amount of collisions.

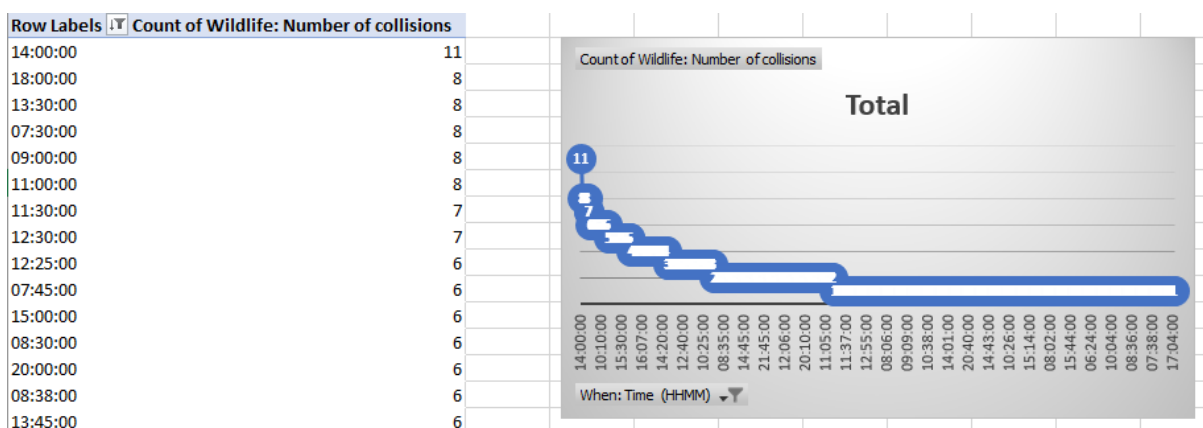


Figure 6.5: Time of day most collisions took place.

Another KPI that was required was the species. Referred to in figure 6.6 is the percentage of species that got hit. There were a high number of gull collisions. This is nearly double of the next closest species which is the rock pigeon.

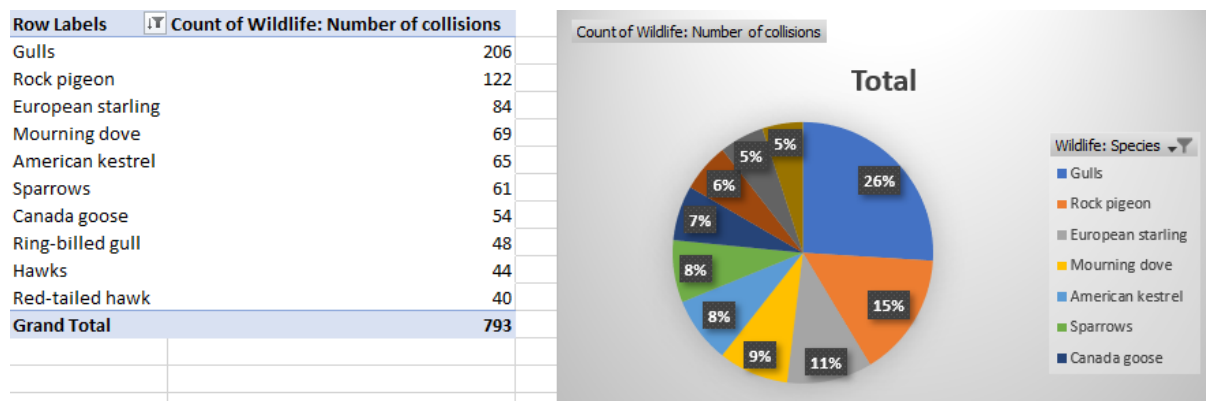


Figure 6.6: Number of collisions per species.

The last pivot table was used to show the number of collisions per aircraft as referred to in figure 6.7. The numbers could be skewed as some aircrafts may have had many more journeys then other aircrafts.

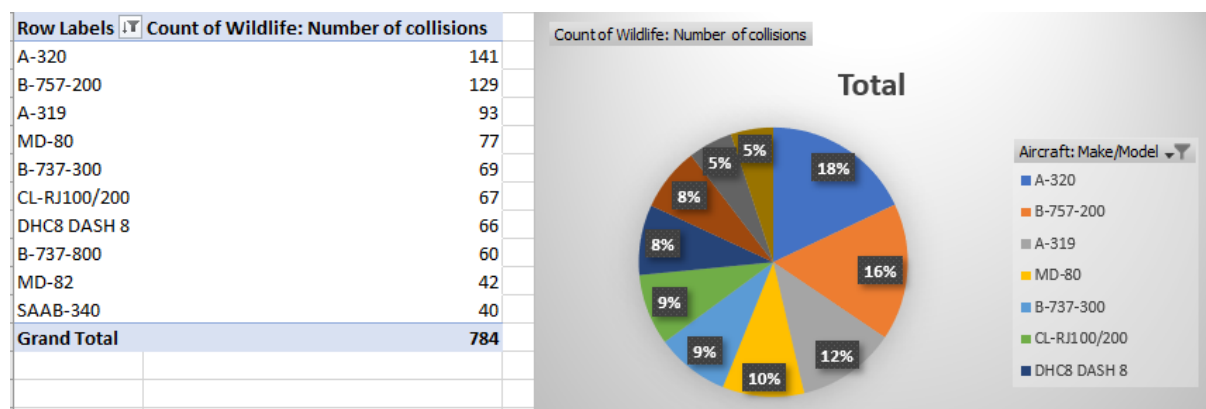


Figure 6.7: Number of collisions per aircraft.

7 Data analysis - preparation and summarisation

After gaining an understanding of the data the next steps is to get the data prepped for importing into QlikView and some data analysis so see if there is some correlation between some of the data. In order to perform the tasks in this section some Excel formulas will be used and to summarize the data filtering will be used. All the numbers seen above will change in the final dashboard in QlikView this is because the data that was previously worked on wasn't fully summarized. As an example, all of the species were still included on this investigation was focusing on all the birds.

This is a study that was authorized by the FAA in America and was only focusing on the top 30 busiest airports in America. Previously the data set contained airports from all over

the world. Found on (wikipedia, 2018) and shown in figure 7.1 is the list of the busiest airports in America and this is what the data set was filtered down to.

Rank (2017)	Airports (large hubs)	IATA Code	Major city served	State	2018	2017 ^[1]	2016 ^[2]	2015 ^[3]	2014 ^[4]	2013 ^[5]	2012 ^[6]	2011 ^[7]	2010 ^[8]	2009 ^[9]
1	Hartsfield–Jackson Atlanta International Airport	ATL	Atlanta	GA		50,251,962	50,501,858	49,340,732	48,004,273	45,308,407	45,798,809	44,414,121	43,130,585	42,280,868
2	Los Angeles International Airport	LAX	Los Angeles	CA		41,232,410	39,636,042	36,351,220	34,314,197	32,425,892	31,326,268	30,528,737	28,857,755	27,439,897
3	O'Hare International Airport	ORD	Chicago	IL		38,593,028	37,589,899	36,305,668	33,686,811	32,317,835	32,171,743	31,892,301	32,171,831	31,135,732
4	Dallas/Fort Worth International Airport	DFW	Dallas/Fort Worth	TX		31,861,933	31,283,579	31,589,832	30,766,940	29,038,128	28,022,877	27,518,358	27,100,656	26,663,984
5	Denver International Airport	DEN	Denver	CO		29,809,091	28,267,394	26,280,043	26,000,591	25,496,885	25,799,832	25,667,499	25,241,962	24,013,069
6	John F. Kennedy International Airport	JFK	New York	NY		29,533,154	29,239,151	27,782,369	26,244,928	25,036,358	24,520,943	23,664,830	22,934,047	22,710,272
7	San Francisco International Airport	SFO	San Francisco	CA		26,800,016	25,707,101	24,190,549	22,766,008	21,704,626	21,284,224	20,038,679	19,359,003	18,467,908
8	McCarran International Airport	LAS	Las Vegas	NV		23,364,185	22,833,267	21,824,231	20,551,016	19,946,179	19,941,173	19,854,759	18,996,738	19,445,952
9	Seattle–Tacoma International Airport	SEA	Seattle/Tacoma	WA		22,639,120	21,887,110	21,231,781	18,781,489	17,450,425	16,625,787	16,425,732	15,408,243	15,273,092
10	Charlotte Douglas International Airport	CLT	Charlotte	NC		22,011,225	21,511,880	21,813,166	21,542,277	21,346,601	20,032,426	19,022,535	18,629,181	18,165,476
11	Newark Liberty International Airport	EWK	New York/Newark/New Jersey	NJ		21,571,194	19,923,009	18,684,818	17,680,826	17,546,596	17,035,098	16,814,092	16,571,754	16,659,441
12	Orlando International Airport	MCO	Orlando	FL		21,565,444	20,283,541	18,759,938	17,278,808	16,884,524	17,159,425	17,250,415	17,017,491	16,371,018
13	Phoenix Sky Harbor International Airport	PHX	Phoenix	AZ		21,185,440	20,896,265	21,351,445	20,344,887	19,525,109	19,556,189	19,750,308	18,907,171	18,559,647
14	Miami International Airport	MIA	Miami	FL		20,709,205	20,875,813	20,986,341	19,468,523	19,420,089	18,987,488	18,342,158	17,017,654	16,187,768
15	George Bush Intercontinental Airport	IAH	Houston	TX		19,603,729	20,062,072	20,595,874	19,772,054	18,952,840	19,038,958	19,306,660	19,528,631	19,290,239
16	Logan International Airport	BOS	Boston	MA	20,431,531	19,145,096	17,749,202	16,290,362	15,425,869	14,810,153	14,293,675	14,171,476	13,561,814	12,566,797
17	Minneapolis–Saint Paul International Airport	MSP	Minneapolis/St. Paul	MN		19,002,544	18,123,844	17,634,252	16,972,678	16,280,835	15,943,751	15,895,653	15,512,487	15,551,206
18	Detroit Metropolitan Airport	DTW	Detroit	MI		17,325,600	16,826,287	16,255,520	15,775,941	15,683,523	15,599,877	15,716,865	15,643,890	15,211,402
19	Fort Lauderdale–Hollywood International Airport	FLL	Fort Lauderdale	FL		16,216,686	14,263,270	13,061,607	11,987,607	11,538,140	11,445,101	11,332,466	10,829,810	10,258,118
20	Philadelphia International Airport	PHL	Philadelphia	PA		14,760,585	14,564,419	15,101,318	14,747,112	14,727,945	14,587,631	14,883,180	14,951,254	15,002,961
21	LaGuardia Airport	LGA	New York	NY		14,737,834	14,706,123	14,762,593	13,415,797	13,372,269	12,818,717	11,989,227	12,001,501	11,084,300
22	Baltimore–Washington International Airport	BWI	Baltimore/Washington, D.C.	MD		13,214,185	12,340,972	11,738,828	11,022,200	11,132,731	11,183,965	11,067,317	10,848,683	10,338,960
23	Salt Lake City International Airport	SLC	Salt Lake City	UT		12,098,835	11,143,738	10,634,519	10,139,085	9,668,048	9,579,836	9,701,756	9,910,493	9,903,821
24	Ronald Reagan Washington National Airport	DCA	Washington, D.C.	VA		11,966,354	11,470,854	11,242,375	10,057,794	9,838,034	9,462,206	9,053,004	8,736,804	8,490,288
25	Washington Dulles International Airport	IAD	Washington, D.C.	VA		11,407,107	10,596,942	10,363,918	10,415,948	10,570,993	10,785,683	11,043,829	11,276,481	11,132,098
26	San Diego International Airport	SAN	San Diego	CA		11,107,078	10,340,164	9,985,739	9,333,152	8,878,772	8,686,592	8,465,683	8,430,509	8,453,854
27	Midway International Airport	MDW	Chicago	IL		10,911,970	11,044,353	11,044,387	10,318,311	9,915,646	9,431,796	9,134,676	8,518,957	8,253,620
28	Tampa International Airport	TPA	Tampa	FL		9,830,583	9,194,994	9,150,414	8,531,561	8,267,752	8,216,153	8,174,194	8,137,222	8,263,294
29	Daniel K. Inouye International Airport	HNL	Honolulu	HI		9,743,989	9,656,340	9,656,340	9,463,000	9,466,995	9,210,270	8,643,494	8,740,077	8,739,389
30	Portland International Airport	PDX	Portland	OR		9,940,866	9,538,472	9,071,154	8,340,234	7,878,760	7,452,603	7,142,620	6,808,486	6,430,119

Figure 7.1: Top 30 busiest Airports in America.

Another filter that was used as stated above was to filter the data by just bird species that got hit by aircrafts. The original dataset contained over 90000 rows of data. Once all the filtering was performed the dataset contain just 972 record which is a significant drop.

The next stage looked at some data analysis. One piece of analysis was to see if there is a correlation between the number of collisions and to time the collision happened. Referred to in figure 7.2 and highlighted in the red is the correlation result. This was found using the Excel function:

=CORREL(A990:A1960,B990:B1960)

As shown in figure 7.2 the two columns that contained time and number of collisions were separated from the rest of the table. The correlation formula takes in two array values and if the result is a +1 the correlation is positive and if it is the minus it's a negative correlation. In this case there was no correlation between the variables. Notice the scatter chart in figure 7.2 the trendline is showing as completely flat referring to the negative correlation.

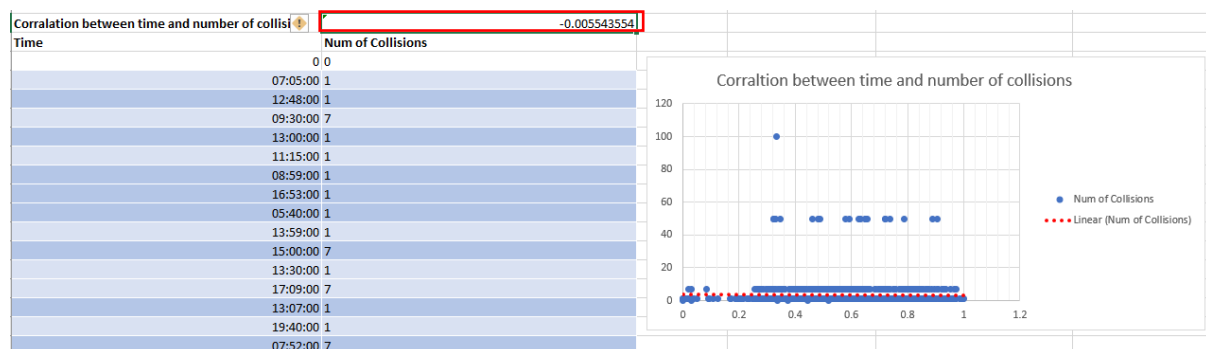


Figure 7.2: Correlation between time and number of collisions.

Another analysis technique used was to find the coefficient of variation. Show in figure 7.3 is two tables the first is the wild life collisions and the second is the collision times. Using Excel functions to get the mean, mode, median and standard deviation you can then work out the coefficient of variation by dividing the standard deviation/mean. The coefficient of variation of wild life collisions is 231 and the coefficient of variation of time is 20.32. Also found was the min, max and range of all those columns.

Wild Life Collisions		Collision Times	
Coefficient of Variation	231.74	Coefficient of Variation	20:32:32
Standard Deviation	7.9904662	Standard Deviation	05:01:23
Mean	3	Median	13:27:00
Median	1	Mean	13:37:45
Mode	1	Mode	14:00:00
Max	100	Max	23:59:00
Min	0	Min	00:01:00
Range	100	range	23:58:00

Figure 7.3: Coefficient of variation

8 Software

Once all the data was cleaned and analysed the next step is load the data into QlikView to do this you must go to file then edit script. Then go to table files and locate the XLS file and click next and you are presented with the screen shown in figure 8.1. Under labels click use embedded labels this option will use the column names instead of the original column names the click finishes and reload the script and all the data is loaded in.

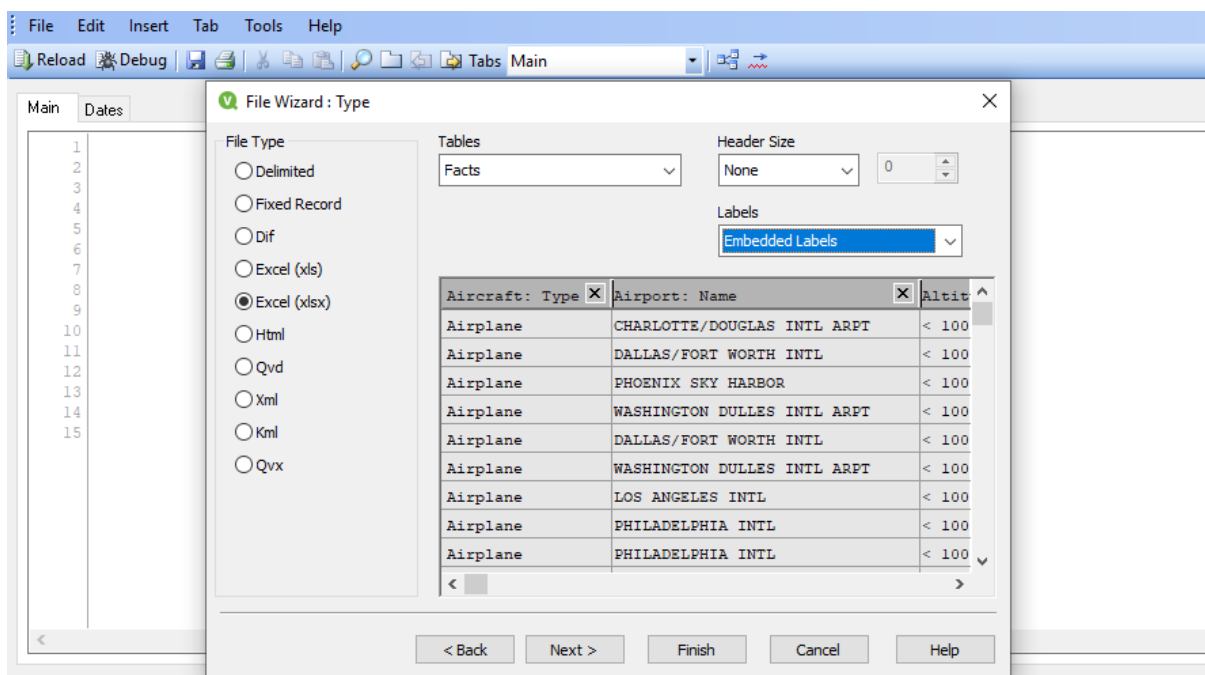
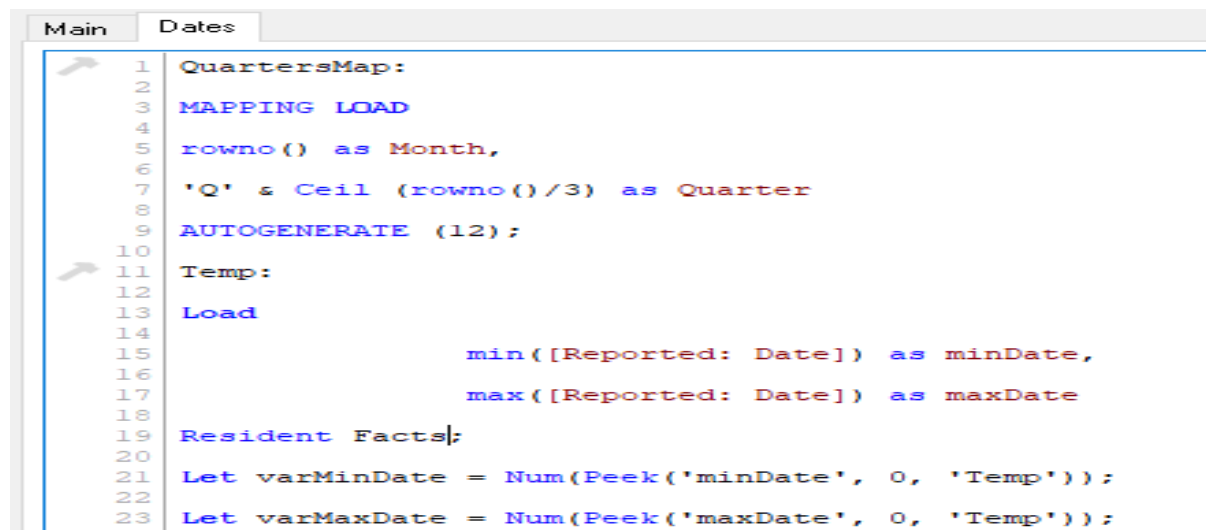


Figure 8.1: Loading XLS script.

Because of the way the date was formatted in Excel MM/DD/YYYY a script was needed format it so it could be separated by years, quarters, months, weeks and day. After some searching a tutorial was located at (community.qlik.com, 2012). This tutorial talks through and explains all the code which was really helpful. Referred to in figure 8.2 is the script named

dates. This script references the table that was imported from Excel called facts. It then formats the date to how you want it.



```

1 QuartersMap:
2
3 MAPPING LOAD
4
5 rowno() as Month,
6
7 'Q' & Ceil (rowno()/3) as Quarter
8
9 AUTOGENERATE (12);
10
11 Temp:
12
13 Load
14
15         min([Reported: Date]) as minDate,
16
17         max([Reported: Date]) as maxDate
18
19 Resident Facts;
20
21 Let varMinDate = Num(Peek('minDate', 0, 'Temp'));
22
23 Let varMaxDate = Num(Peek('maxDate', 0, 'Temp'));

```

Figure 8.2: Master calendar.

Once the data is loaded it was then time to build the dashboards. Using Excel to understand the data has major benefits because as well as understanding you create the charts. By doing so it was just a matter of replicating the charts in QlikView. Using the chart wizard or by right clicking on the dashboard and choosing new sheet object it is relatively straight forward. In the first few drafts there was a lot of use of pie charts and they weren't suitable for what was needed. A switch to bar chart and by reducing the number of dimensions shown in them made the charts far more readable. Referred to in figure 8.3 is the dashboard that contains all information about the aerospace collisions. The dashboard navigation to the left is a container object with a search, calendar and current selection object inside it. Users may filter by the calendar date and when a chart is clicked it will populate to the current selection box.

After that there is two horizontal bar charts representing the top ten airports and airlines for collisions. There is a pie chart shown to collisions by aircraft model and a table that shown the rest of the data got to do with aerospace. If a user clicks on the others in the charts they can view more information.

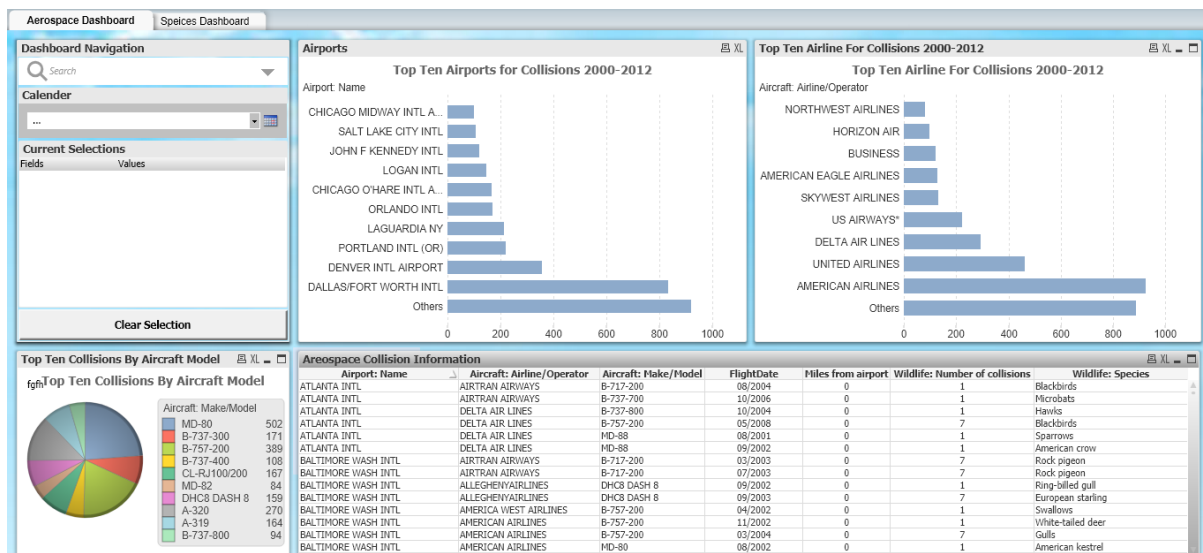


Figure 8.3: Aerospace dashboard.

Shown in figure 8.4 is all information containing collisions with the species. The dashboard navigation is the same object as in the aerospace dashboard. After that there is a vertical bar chart that shows the total collision for each quarter. There is a drill down where users can drill down to year and month. The chart beside that is a line chart the shows at what part of the day the most collisions are happening. There are two pie charts one represents the total species that get hit and the second shown the time most collisions happen. All charts are linked by clicking on one will populate the rest. Finally, at the bottom is a straight table shown all the data related to the species and collisions.

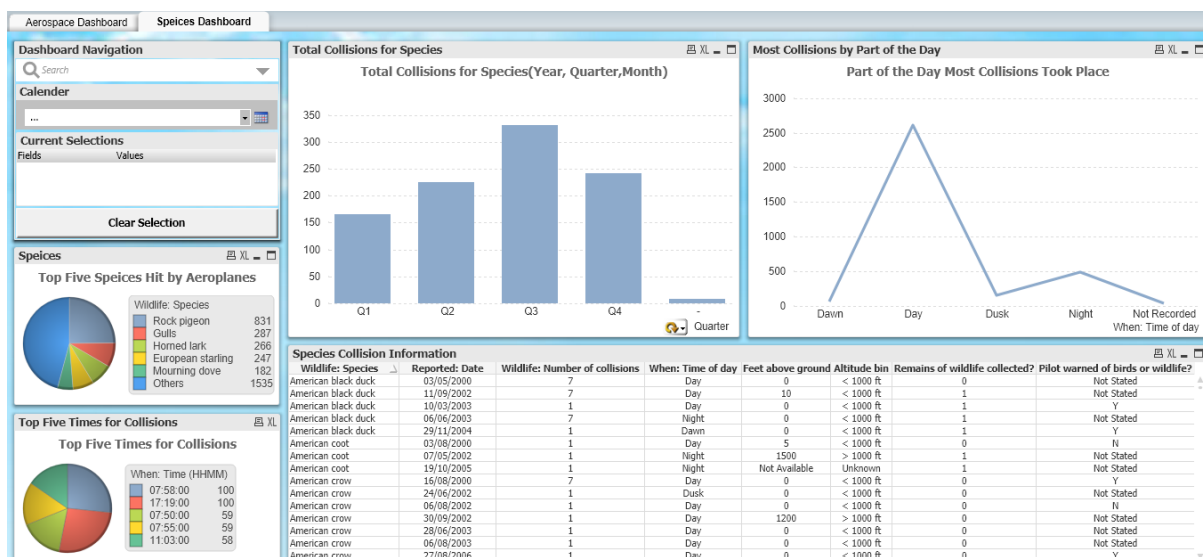


Figure 8.4: Species dashboard.

9 Limitations

Unfortunately, the limitation doesn't lie with the software, it lies with the designer. QlikView makes it quite simple to build a straight forward dashboard. But with experience and time you could build a dashboard with greater detail. When building this dashboard while searching through the QlikView help pages you could see some really advanced functions that

are used but this is a course on its own and you would need time to learn how to use this software properly.

10 Conclusion

That concludes this report for the FAA Investigation into Wildlife Collisions 2000-2012. The aim of this report was to demonstrate the tools that would be used by an industry to clean, analyse, understand, represent and present data. Excel was used to demonstrate how to clean, understand, analyse and represent the data. QlikView was used to present the data. By doing this report it gives us a good understanding on how powerful Excel is for data analytics. QlikView makes building nice dashboards very simple through its graphical interface.

11 References

community.qlik.com, 2012. *community.qlik.com/t5/QlikView-Scripting/Creating-A-Master-Calendar*. [Online]

Available at: <https://community.qlik.com/t5/QlikView-Scripting/Creating-A-Master-Calendar/td-p/341286>

[Accessed 29 03 2019].

FAA, 2019. *www.faa.gov*. [Online]

Available at: <https://www.faa.gov/about/mission/activities/>

[Accessed 25 03 2019].

wikipedia, 2018. *en.wikipedia.org*. [Online]

Available at: https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States

[Accessed 29 03 2019].

YouTube, 2016. *Excel Magic Trick 1262: Convert Times YouTube*. [Online]

Available at: <https://youtu.be/-Rkk4EpXskc>

[Accessed 25 03 2019].