ADVANCED BUSINESS ANALYTICS : END TERM

NAME: SAURABH SHARMA, ROLL NO: 190041

SECTION A

QUES 1 : What are the appropriate measures (numerical variables)?

ANS:

The dataset has a collection of Categorical and Continuous variables that is numerical variables that collectively constitutes the data information.

It has the following numerical measures :

- **Record ID** that shows the records of Bird Strikes happened with a random ID.
- **Wildlife: Number Struck Actual**: that shows the actual number of Wildlife that stroked with the type of Aircraft.
- **Aircraft: Number of engines** - shows the number of engines the aircraft was having.
- **Cost: Total $** shows the cost incurred for the damage happened to aircraft.
- **Feet above ground** shows the altitude when the bird strikes the aircraft.
- **Number of people** injured shows the people injured due to that accident.

QUES 2 : Statistics related to "feet above ground" mention Mean, Median, Min, Max

ANS:

For variable Feet Above Ground :

Mean lies around : 795.7316

Median around: 50

Min around: 0

Max around: 18000

Statistics	
Mean	795.7316
Median	50
Min	0
Max	18000
Standard Deviation	1740.7245
Unique Values	236
Missing Values	0
Feature Type	Numeric Feature

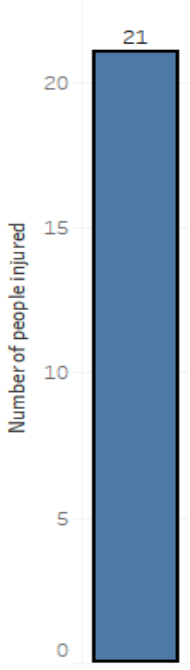
That indicated that the values for feet above ground starts from 0 and goes to 18000 Feet above ground. Zero Feet here may indicate that aircraft was may be about to land or about to take-off.

Highest altitude that any wildlife has encountered to the aircraft is 18000 ft. Where as the average for all the strikes lies around 795.73 ft, after doing the missing values treatment.

QUES 3 : How many total human fatalities are mentioned in the dataset?

ANS:

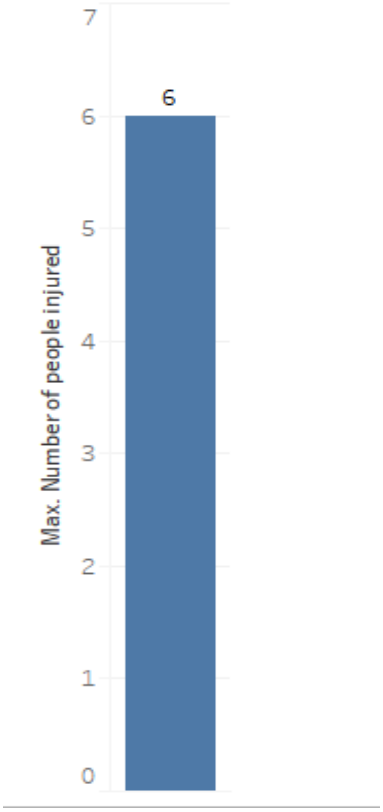
Summing all the values incorporated through variable **Number of people injured, there are 21 humans** who were injured with this accident.



QUES 4: Maximum number of humans injured in one single incidence.

ANS:

Maximum number of people that injured in one-time single incident are **6 People.**



QUES 5: Average speed (in knots) when the accidents happen

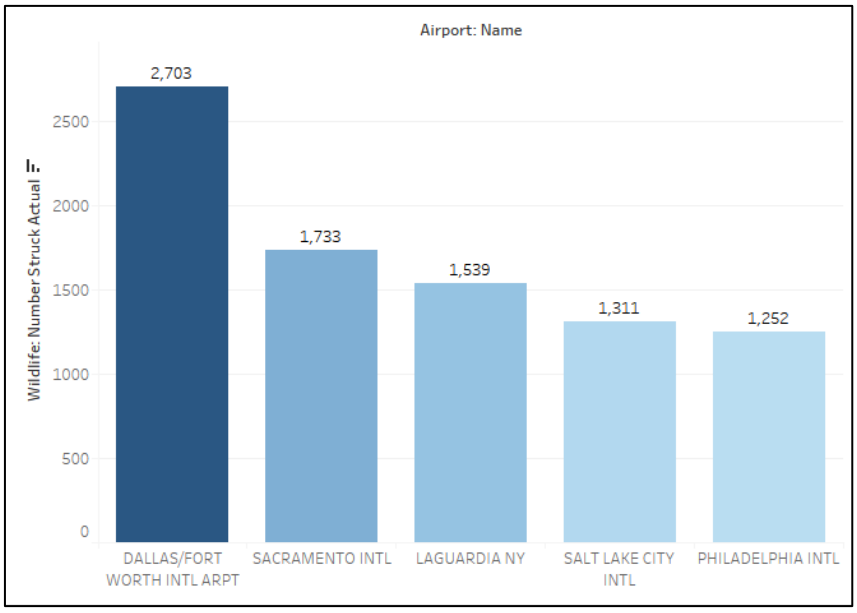
ANS ***No such information or relevant data is given to calculate the speed when accidents happened.

QUES 6 : Top 5 airports known for wild life incidents.

ANS:

TOP 5 Airport known for Wild Life Incidents are as follows:

- DALLAS/FORT WORTH INTL ARPT
- SACREMENTO
- LAGUARDIA
- SALT LAKE CITY
- PHILADELPHIA INTL ARPT

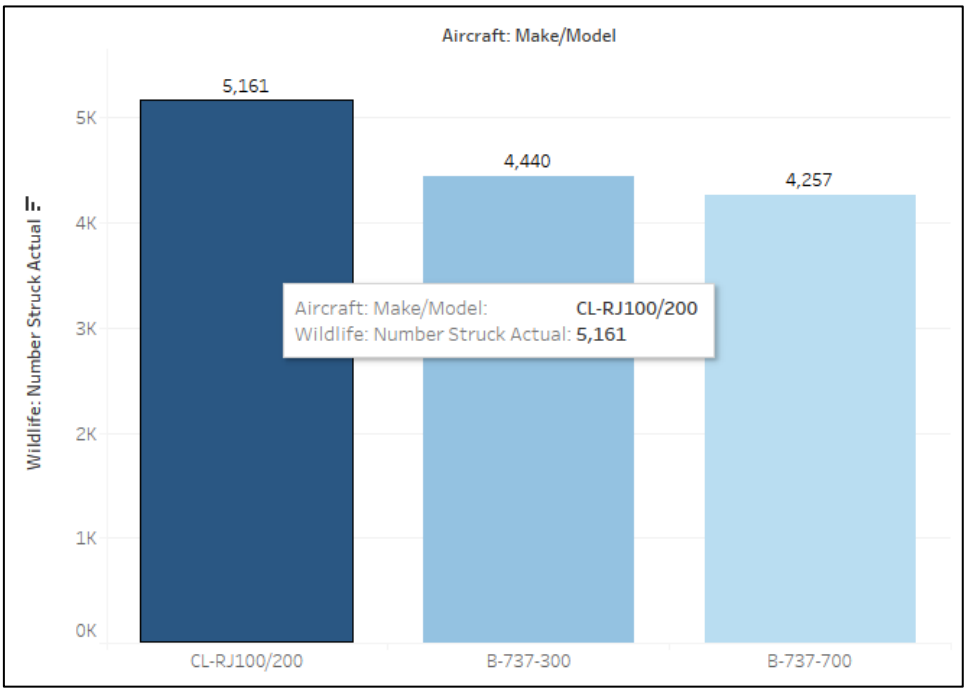


QUES 7: Top 3 aircraft makes for wild life incidents.

ANS

The top 3 Make/Models that were stroked with the accident are:

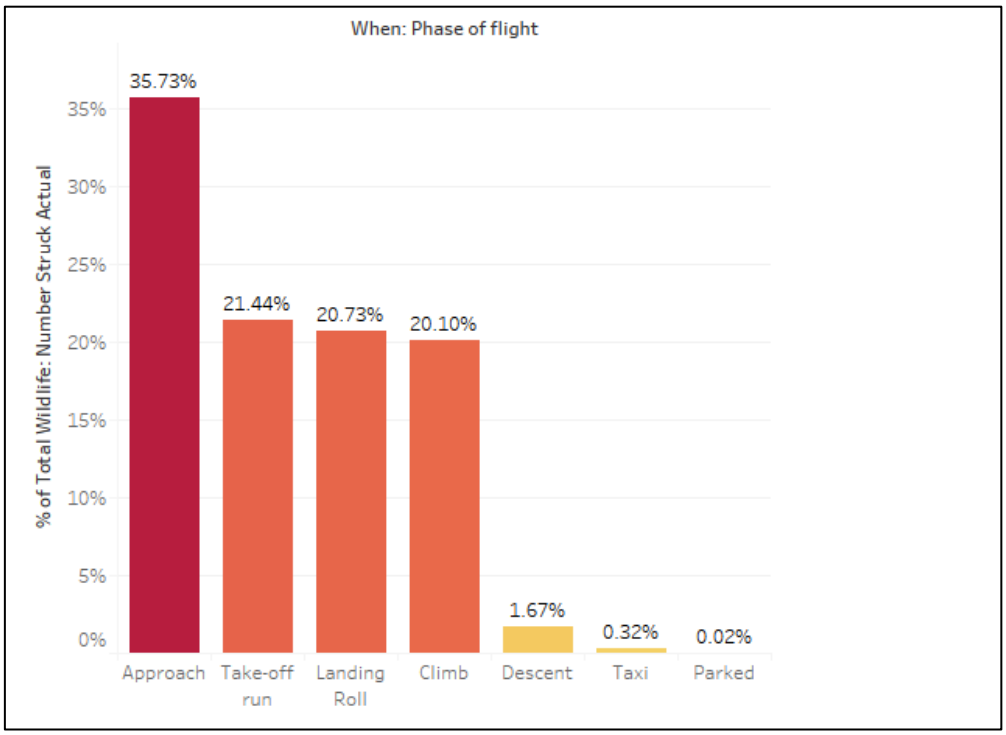
1. **CL-RJ100/200**
2. **B-737-300**
3. **B-737-700**



QUES 8: What is the percentage wise distribution of incidents on the basis of phase of the flight?

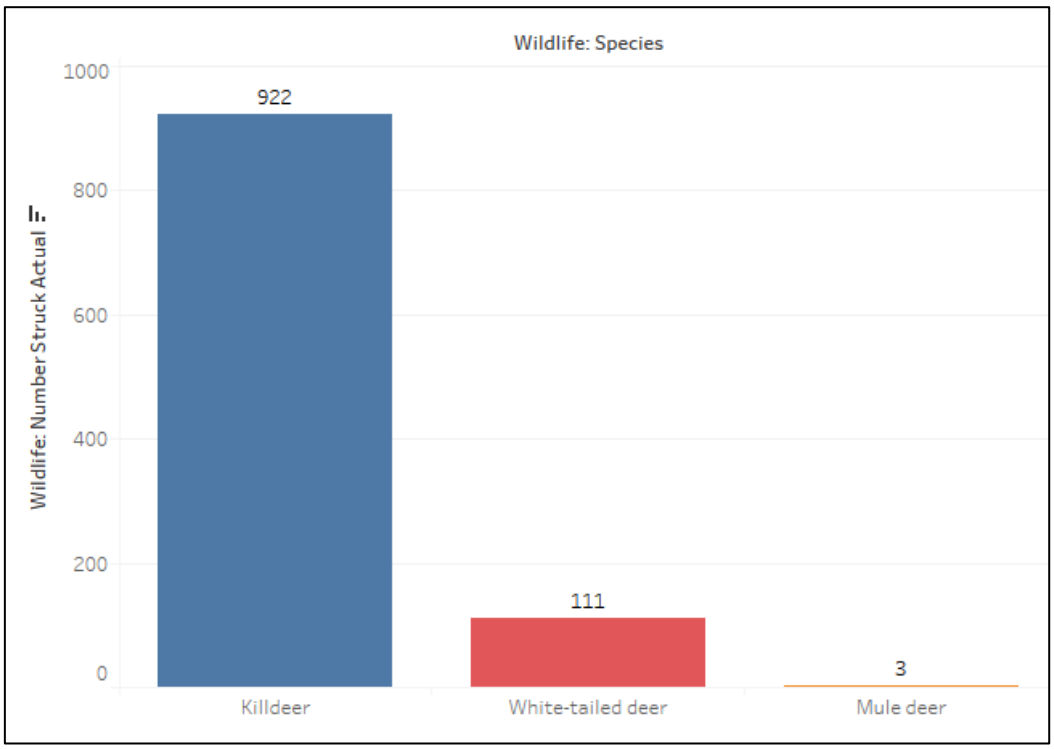
ANS:

- Approach Phase contributed to around 35.73% to total incidents happened and is the highest.
- Followed by Take-off Phase, Landing Roll and climb which are almost neighbours.
- Parked attracted least number of incidents that is around 0.02%.

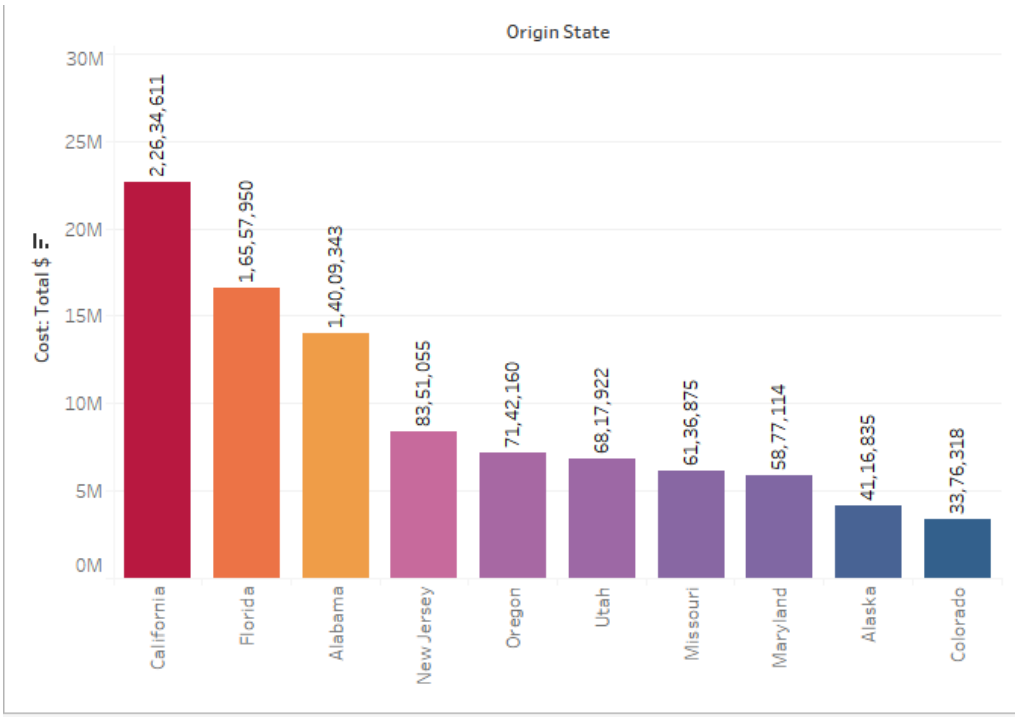


QUES 9: How many times "deers" were involved?
ANS:
There was total 3 species for Deers found namely-

- **Killdeer contributed 922 incidents.**
- **White-tailed deer contributed 111 incidents. And**
- **Mule Deer contributed the least i.e., 3 incidents**



QUES 10: In which state in the US the maximum cost was incurred for wild life strike?
ANS
According to data given, California incurred the highest cost i.e., around $ 2,26,34,611



SECTION – B

QUES : Perform the text analytics of the column "Remarks".

ANS. Text analytics is the automated process of translating large volumes of unstructured text into quantitative data to uncover insights, trends, and patterns. Using Azure ML Text Analytics, there are some techniques through which I have tried to analyse the remarks variable.

Techniques Used :

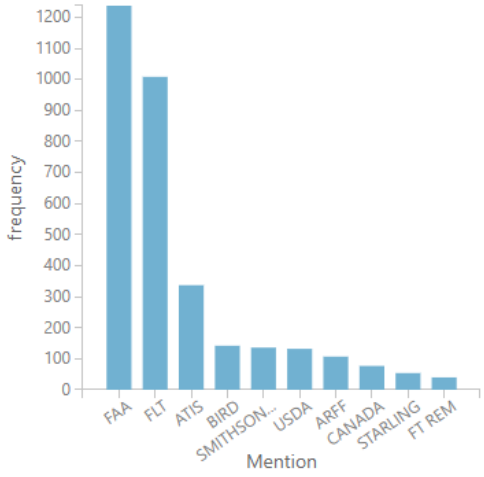
1. Language Detection.
2. Named Entity Recognition.
3. Entity Key Phrases from text.
4. Latent Dirichlet Allocation**.**
5. Extraction of N-grams Features
6. Checking the Model Accuracy.

Summary :

1. Language English is used almost for giving remarks.
2. For the entities : Highest used Entities are

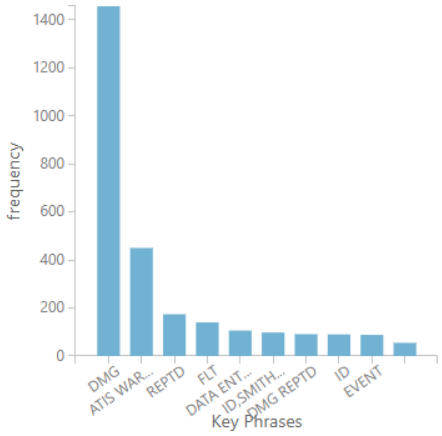
- FAA
- FLT
- ATIS
- BIRD
- SMITHSONIAN
- USDA
- ARFF

And so on.

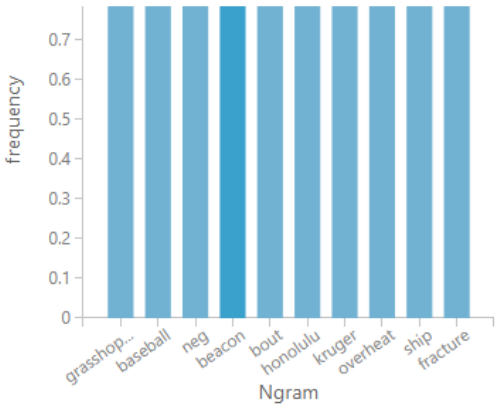


3. Some of the Key Phrases extracted out of the remarks are :

- DMG
- ATIS WARNING
- REPTD
- FLT
- DATA ENTRY NOTE REPTD
- ID SMITHSONIAN
- EVENT



4. In the N-gram Extractions, various tokens created and extracted out through we can easily summary out that the remarks are about some birds, some different species, overheating of aircraft , injuries like fractures are remarked by the respondents.



SECTION – C

QUES 3:

What other machine learning algorithms are possible to be used in this data? You only need to describe the possibilities and it is not necessary to build the model.

ANS:

After having the data in hand, I can see the missing values first in the data because that can give us distracting outputs. By treating the missing values whether by multiple imputations method or removing the null rows, I will move further. Will check for multi-collinearity issues as well.

Further analysis of the data set has given me a brief understanding about the data. Because it is important to find out the Dependent Variable and Independent variables to perform several modelling techniques on the data.

For ex. I have interpreted the Variable : Effect: Indicated Damage

This shows the Effect of incidents happened, whether it caused any damage or not.

Since this is a Categorical Variable and is dependent on several other variables which are independent in nature, I have used Classification techniques to analyse how this variable performs.

After creating models using classification techniques, we can actually see how accurate the model is working that means how smooth this variable is performing with the set of independent variables.

Techniques I have used :

- Two-Class Logistic Regression : **ACCURACY 91.2% Towards No Damage**
- Two-Class Support Vector Machine : **ACCURACY 90.4% Towards No Damage**
- Two-Class Boosted Decision Tree : **ACCURACY 92.1% Towards No Damage**
- Two-Class Decision Forest : **ACCURACY 88.7% Towards No Damage**

Out of these different models that I have prepared, I got the model accuracies.
So, whichever model gives us the highest model will be best fit for this classification problem.
And the accuracy scores for these models gives the best result through Two-Class Boosted Decision Tree, which I have considered the best fit model in this case. We can perform Hyper Tuning to increase our effectiveness of accuracy

As far as possibilities are concerned, we can use Association rules to analyse some pairs about when there can be probabilities when we had great damages.
After doing the Association I got around – 1177 rules with support 0.4 and confidence 0.7.
Further analysing the pairs, we can identify the probabilities of No damage and Damage.

For Further References:
https://gallery.cortanaintelligence.com/Experiment/BIRD-STRIKE-ANALYSIS
https://gallery.cortanaintelligence.com/Experiment/BIRD-STRIKE-REMARK-TEXT-ANALYSIS

Thankyou.