**CSE3020**

**Data Visualization**



**Visualizing Data for SILK AIRLINES**

**Submitted to:**

**Prof. Tulasi Prasad**

**Group Members:**

**B SAI MOUNIKA-15BCE1099**

**Aim:**

This project mainly aims at analysing the different fields in the given dataset and visualising it with Tableau. And here we are calculating Training accuracy, testing accuracy, and cross-validation accuracy to check the fit of the model by using 5 models Logistic Regression,Support Vector Machines,Random Forests,K-Nearest Neighbours,Gaussian Naive Bayes.Our aim here is to make management, comparison and visualization of the dataset of the airlines, Silk Air, more effective, simpler and meaningful. We will be to see connections more effectively as they occur between operating conditions and performance.

**Dataset & Field Description:**

A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. A data set is organized into some type of data structure.

The dataset that we have chosen corresponds to the data of an airline. We have given it our name, Silk Air Airlines. It contains various fields which not only tell us about the type of aircraft, altitude, speed etc but also the other aspects of airlines. The dataset that we have used gives us an apt description of the aircraft used but also the climatic conditions (eg precipitation, sky) under which it travels. We also get to know the locations, the different kinds of aircraft travel to, and the effects it causes. Also the record ID for each aircraft is provided which gives us all the details about the date and time of arrival and departure of the aircraft. Our dataset comprises of the fields given below:

1. Aircraft Type
2. Airport Name
3. Altitute Bin
4. Aircraft: Make/Model
5. Wildlife: Number struck
6. Effect: Impact to flight
7. Effect: Other
8. Location: Nearby if en route
9. Aircraft: Flight Number
10. FlightDate
11. Record ID
12. Effect: Indicated Damage
13. Location: Freeform en route
14. Aircraft: Number of engines?
15. Aircraft: Airline/Operator
16. Origin State
17. When: Phase of flight
18. Conditions: Precipitation
19. Reported: Date
20. Conditions: Sky
21. When: Time (HHMM)
22. When: Time of day
23. Cost: Aircraft time out of service (hours)
24. Cost: Other (inflation adj)
25. Cost: Repair (inflation adj)
26. Cost: Total $
27. Miles from airport
28. Feet above ground
29. Speed (IAS) in knots

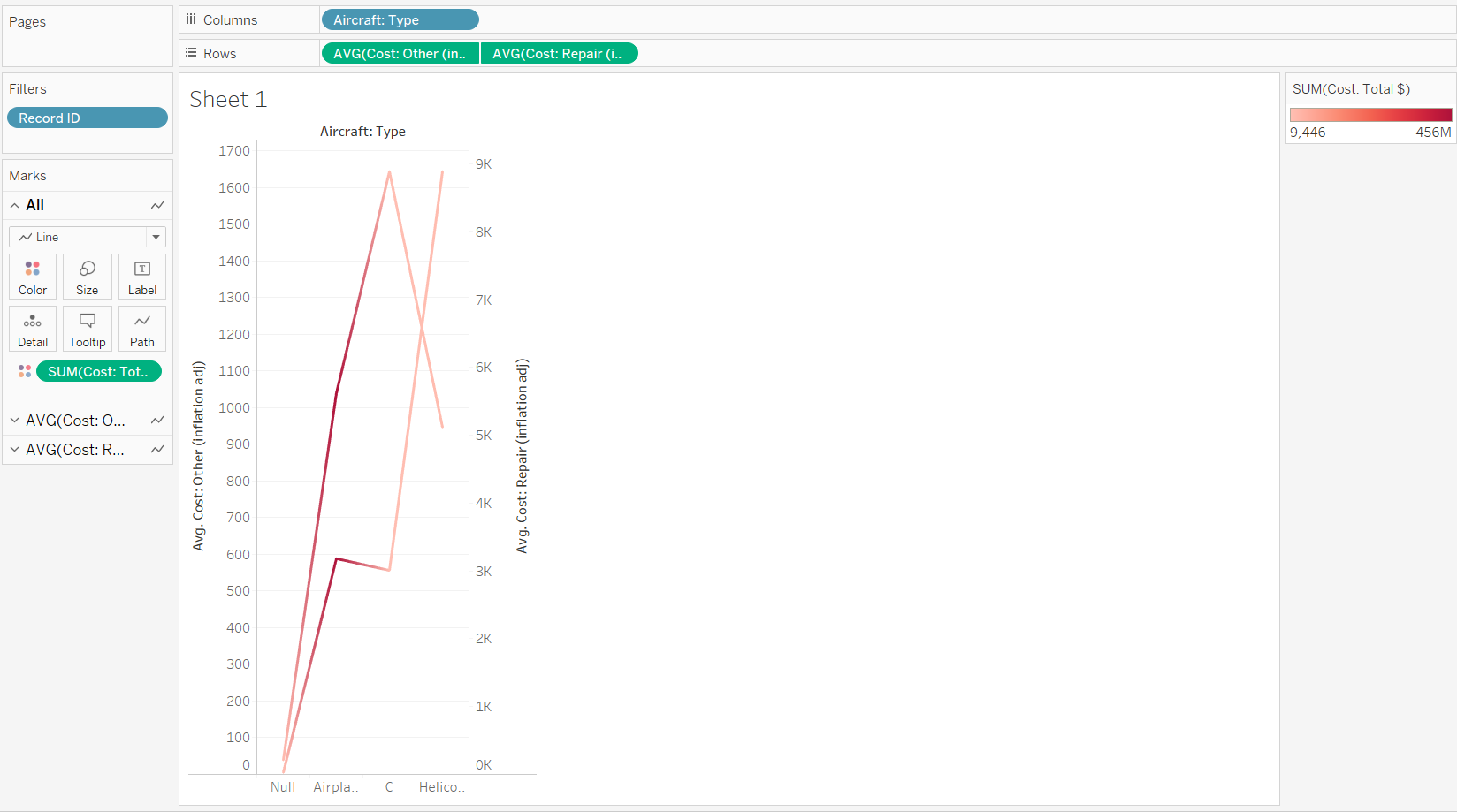
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Data** |  | | **Aircraft Type** | **String** | | **Aircraft Name** | **String** | | **Cost** | **Integer** | | **Effects** | **String** | | **Date** | **Date** | | **Time** | **Timestamp** | | **Origin State** | **String** | | **Distance(miles)** | **Integer** | | **Height** | **Integer** | |  |

**What did we visualize for end user(s)?**

Each sheet describes each aspect of the aircraft and the comparisons made with it attributes. We can derive various conclusions from the sheets which gives us a broader outlook and helps us make a detailed analysis of the data given.

**Sheets Explanation (in Tableau Software):**

Sheet 1:

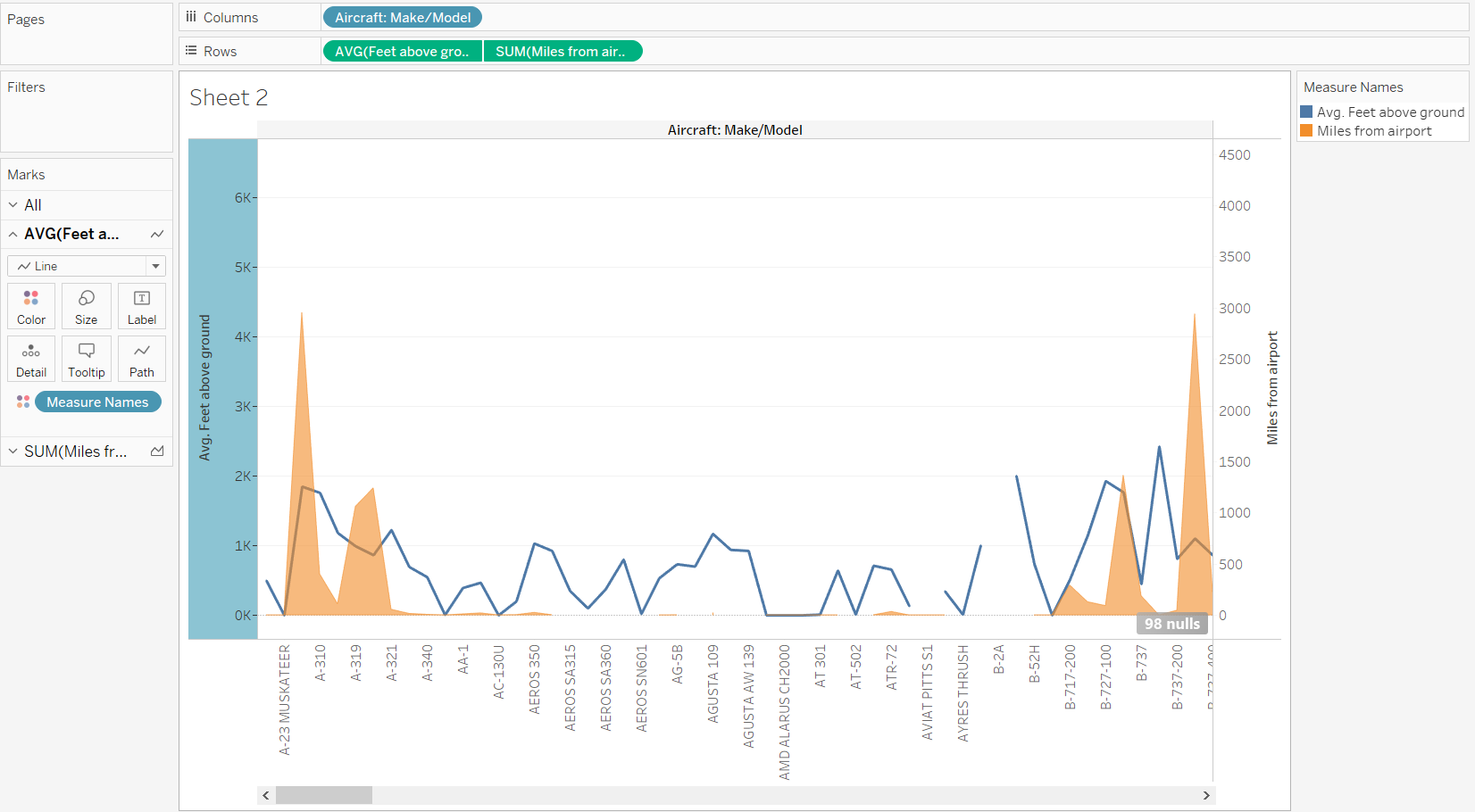


This sheet visualizes different kinds of cost (repair etc.) for an aircraft.

Inflation and Repairing cost of an aircraft type can be visualized and this can be very useful for:

1. Engineers
2. Airlines Finance Department

Sheet 2:



This sheet visualizes the location of a specific aircraft model by visualizing its:

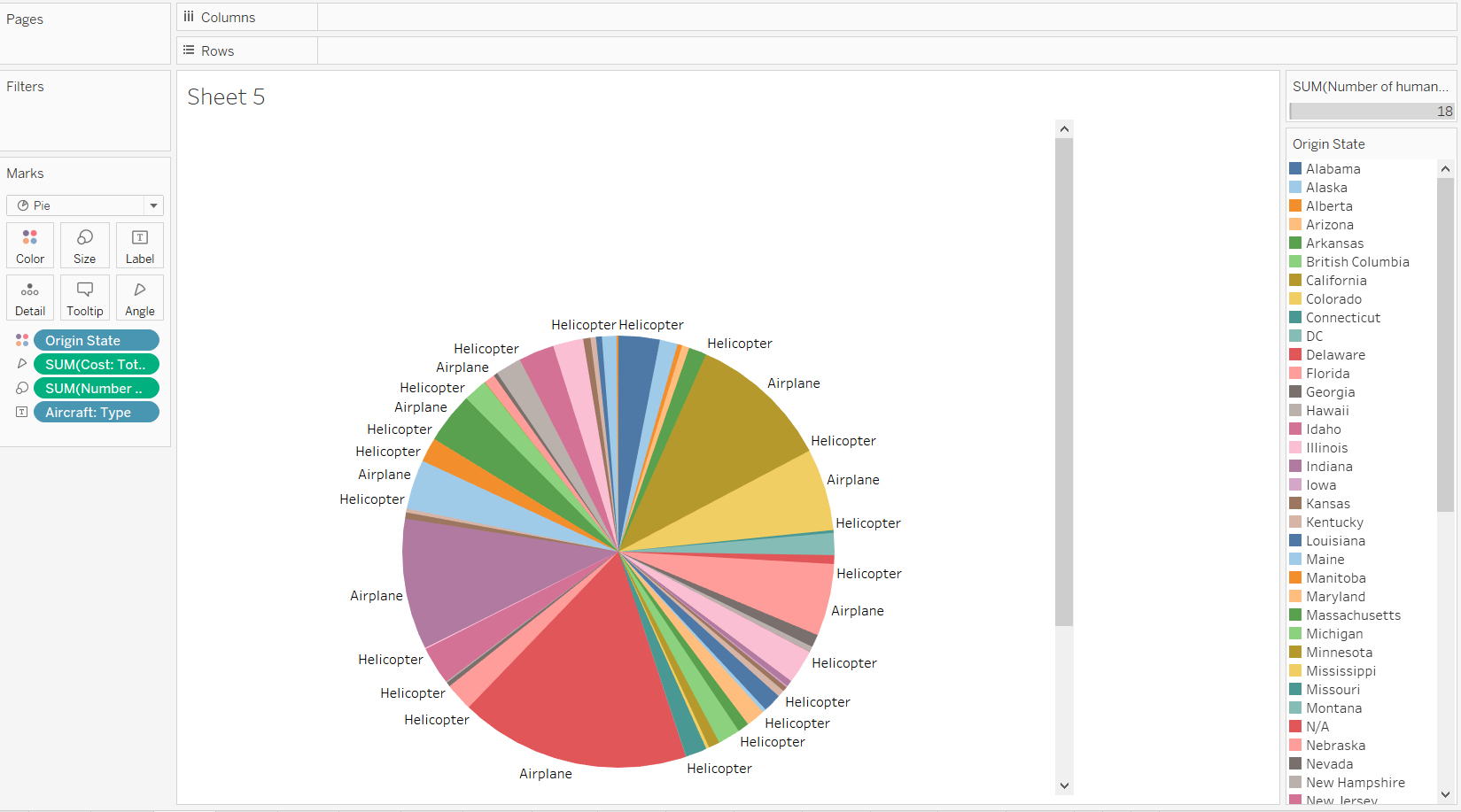
1. Miles from airport
2. And average feet above the ground

This will help the

1. Pilots
2. Engineers and scientists

In determining the location of an aircraft and predict other features

Sheet 3:

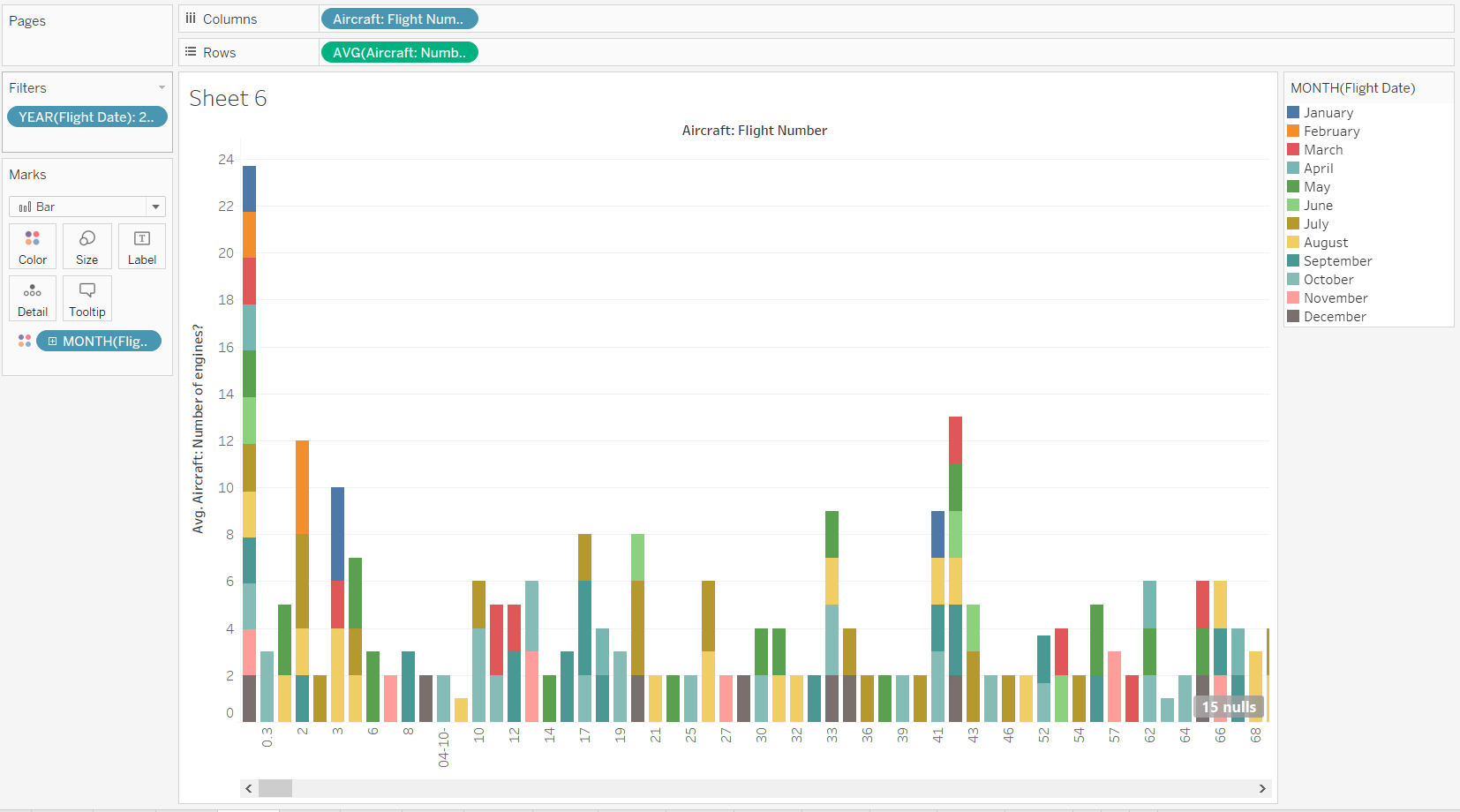


This sheet visualizes different types of aircrafts(origin statewise) and total cost of them in a pie chart, so that in one look we can determine cost of aircrafts coming from different places.

This will help

1. The finance deoartment of airlines
2. Engineers and scientists
3. Manufacturers

Sheet 4:



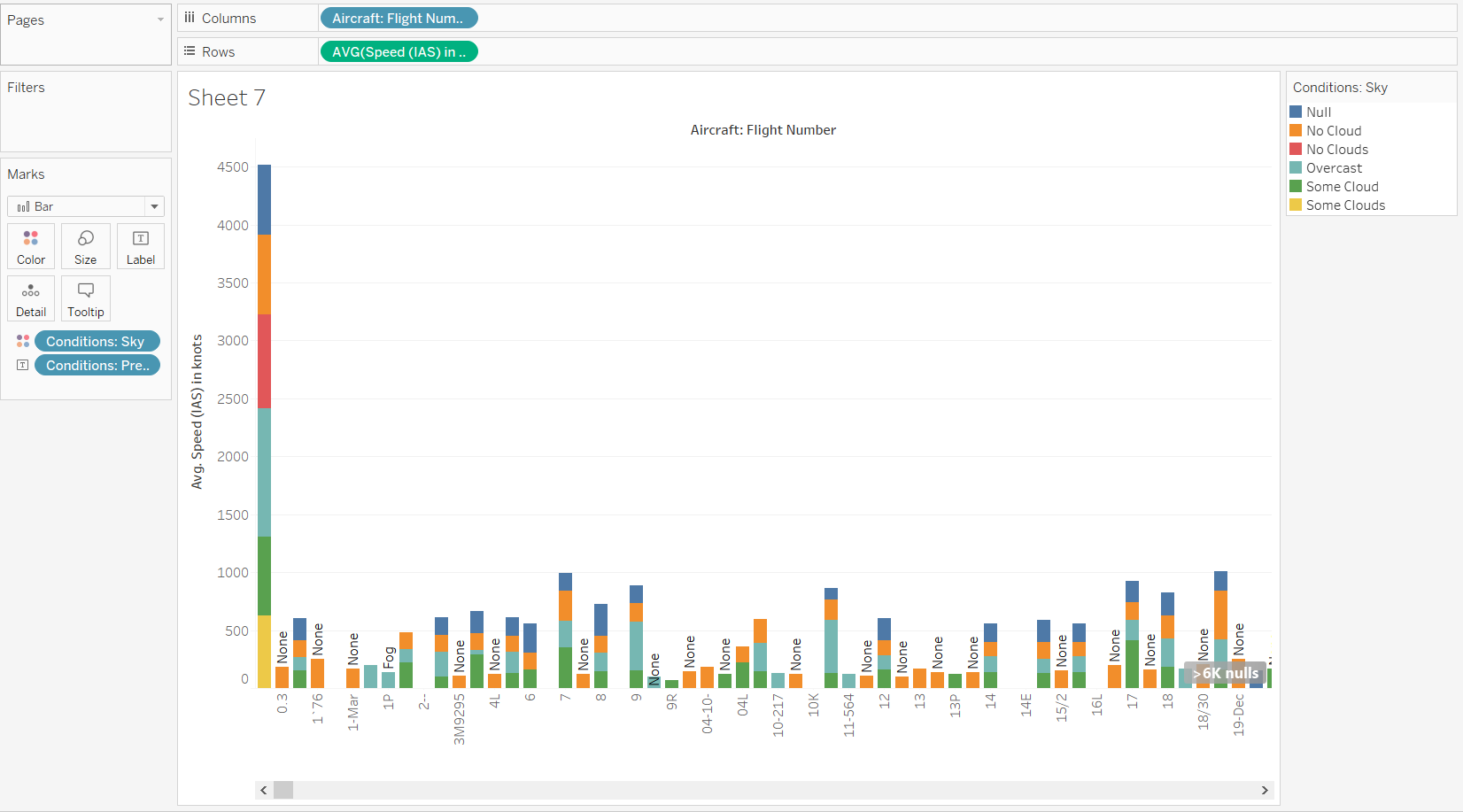
This sheet visualizes Flight details like: Flight Number, Flight Date, Flight time of take-off and number of engines.

This will help:

1. Passengers
2. Airlines staff
3. Pilots
4. Engineers( because they have to be ready for take off)

These are the basic flight details visualized in one bar chart.

Sheet 5:

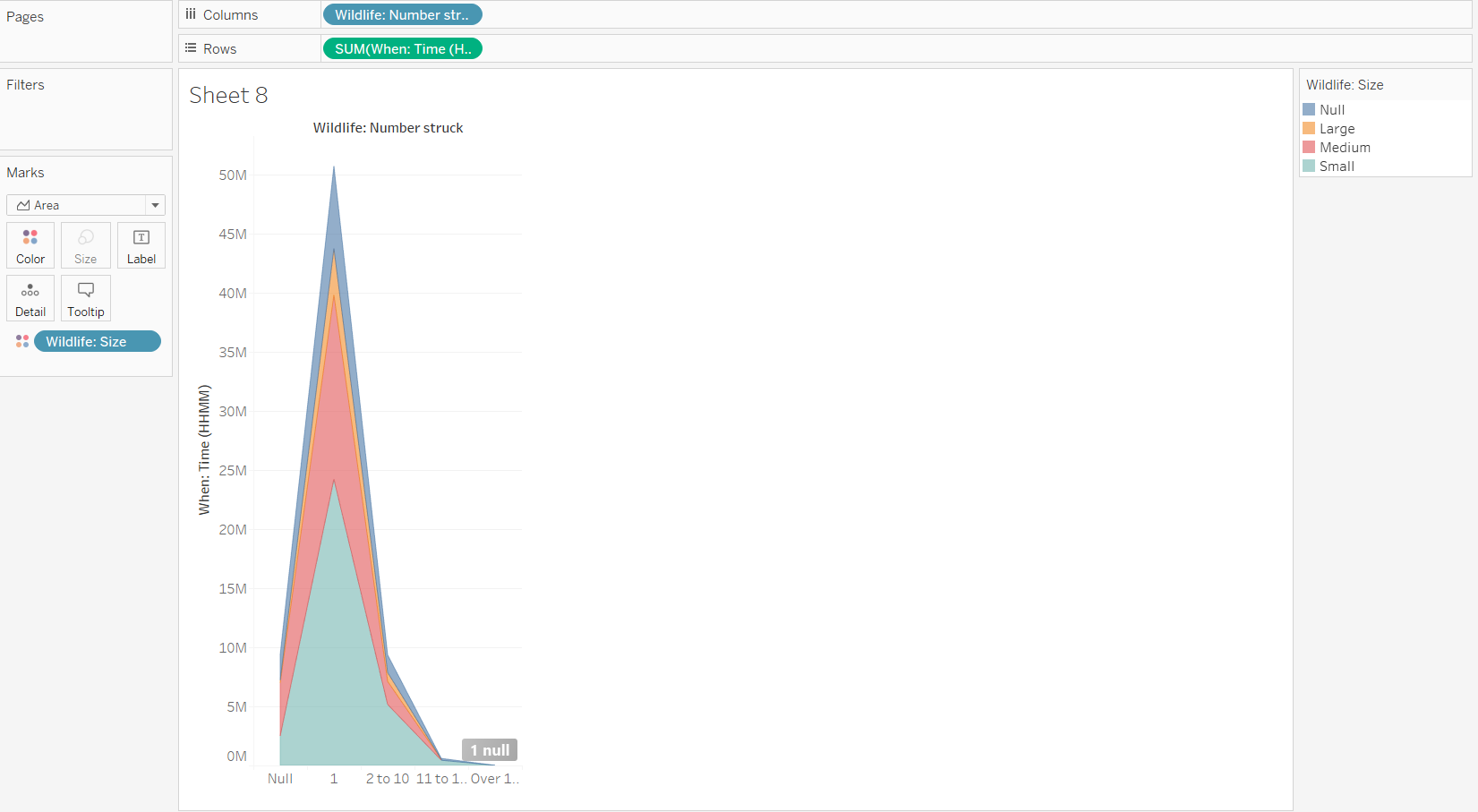


This sheet visualizes Climatic conditions and flight number.

This will help:

1. Pilots
2. Passengers
3. Staff
4. Engineers and Scientists
5. Weather CareTakers

Sheet 6:

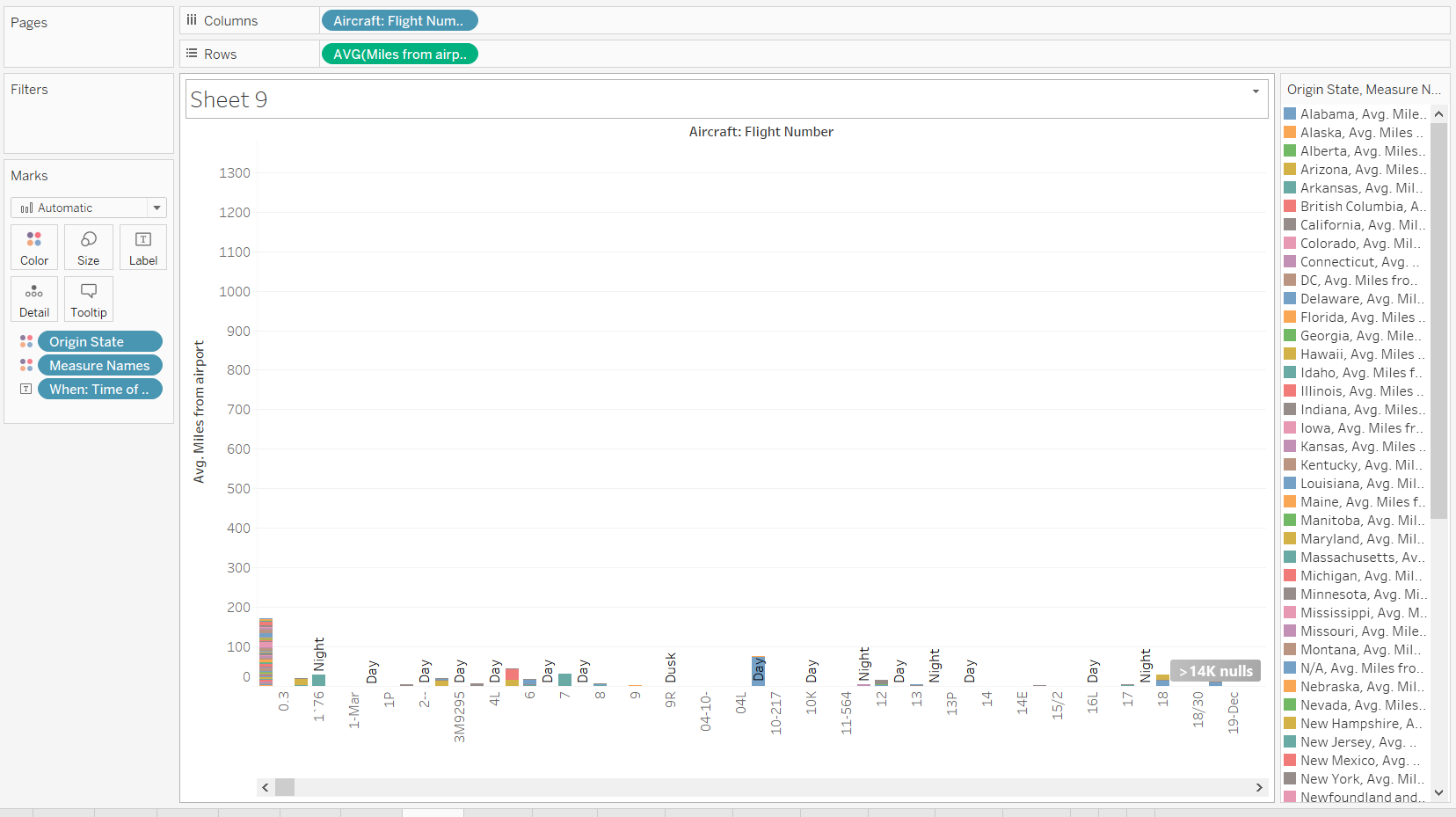


This sheet visualizes the wildlife happenings. The attributes used here are: Number of wildlife struck, Wildlife size, and at what time.

This will help:

1. Engineers and experts- to change the altitude of the flight
2. Pilots- to be careful of the altitude

Sheet 7:



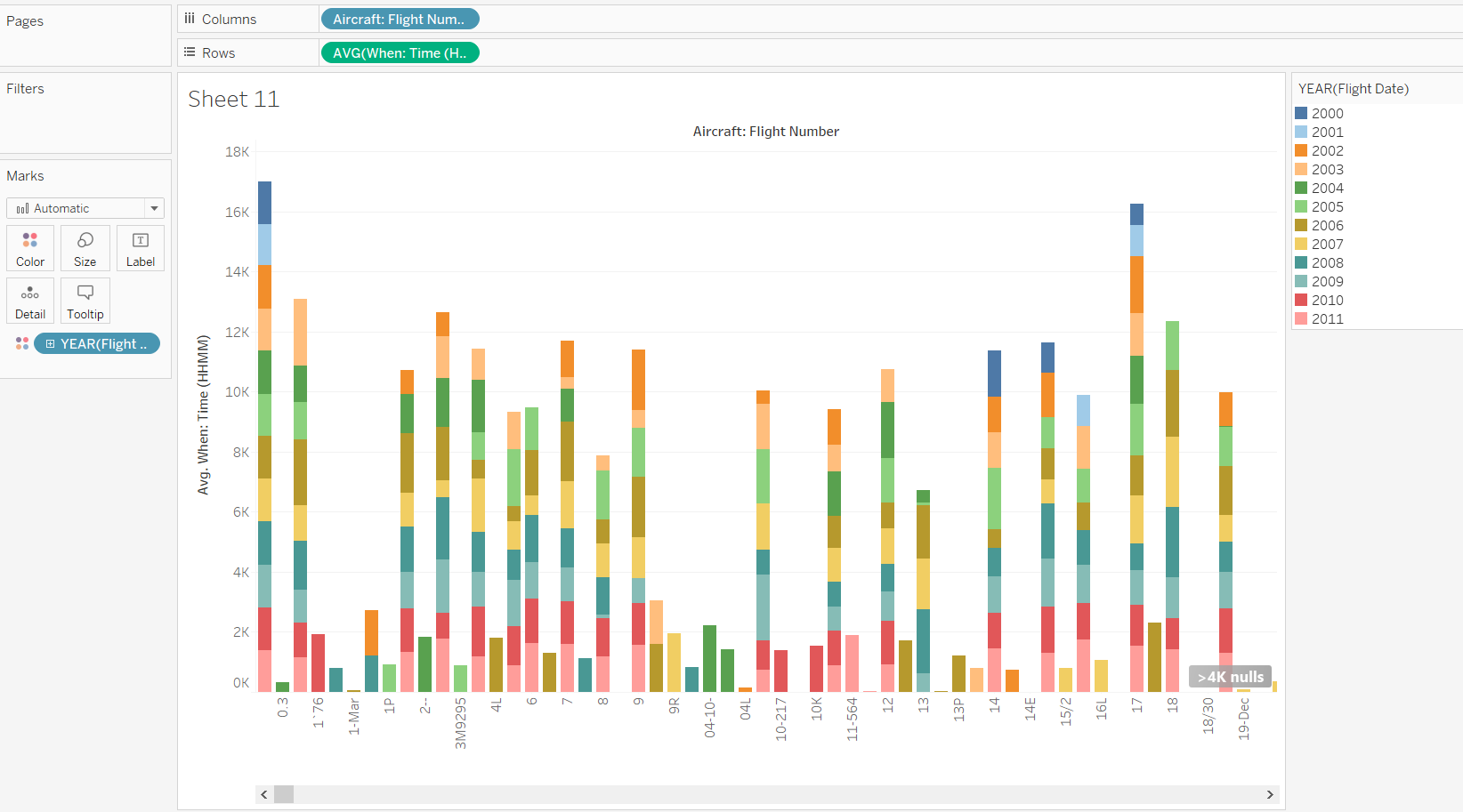
This sheet visualizes the flight number and its location at every point of time.

The attributes like Origin state, time and miles from airport are taken.

This will help:

1. Engineers and Scientists
2. Pilots

To keep track of the distance and time.

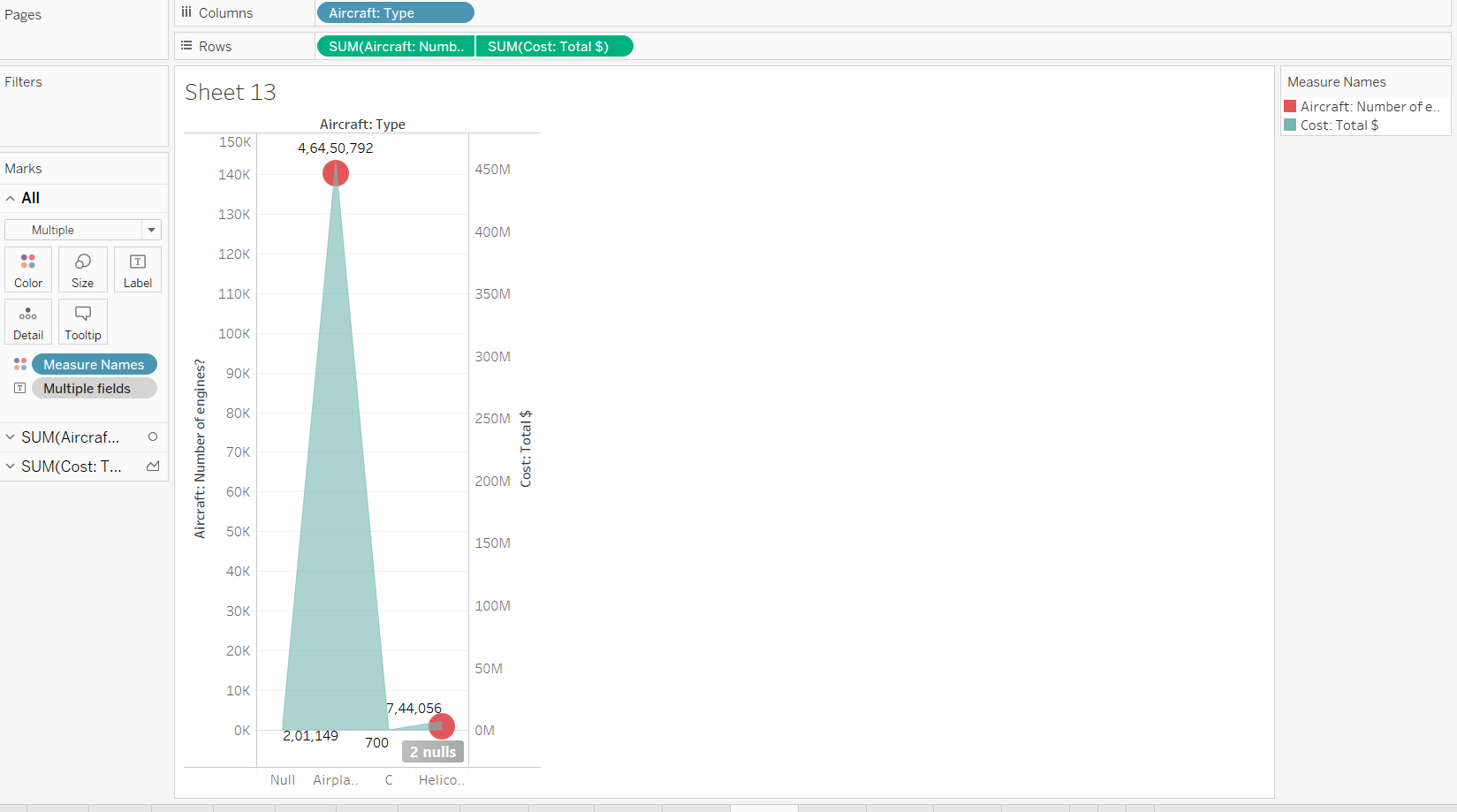
Sheet 8:

This sheet visualizes Flight date and time exclusively for passengers.

This contains basic features like – Fligh number, Date and Time of take off.

This will help – Passenegers.

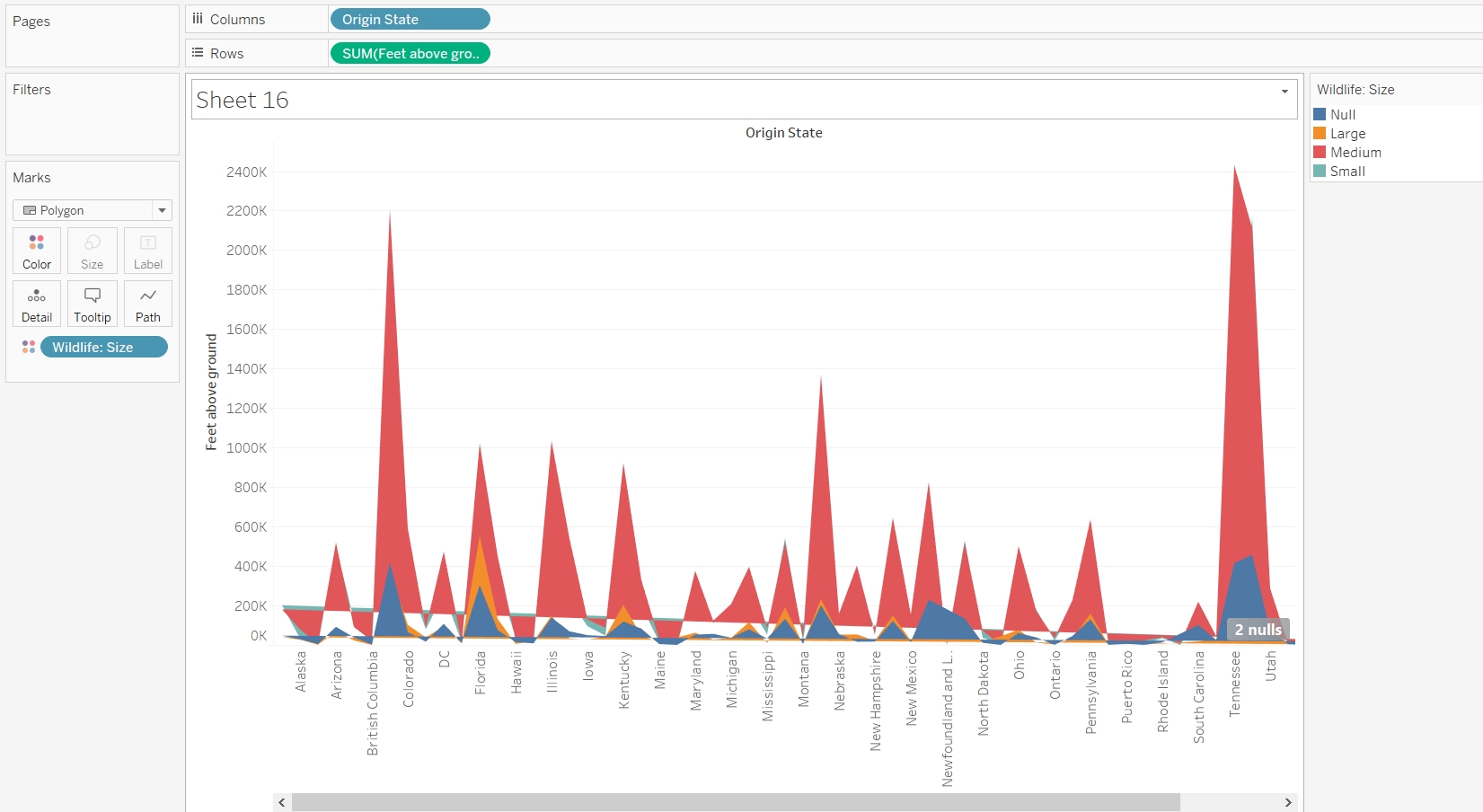
Sheet 9:



This sheet visualizes the aircraft type, their cost and the number of engines used for th aircraft This will help:

1. Different airlines to keep track of the number of engines which are to be used.
2. End Users are- Engineers and Scientists
3. Pilots

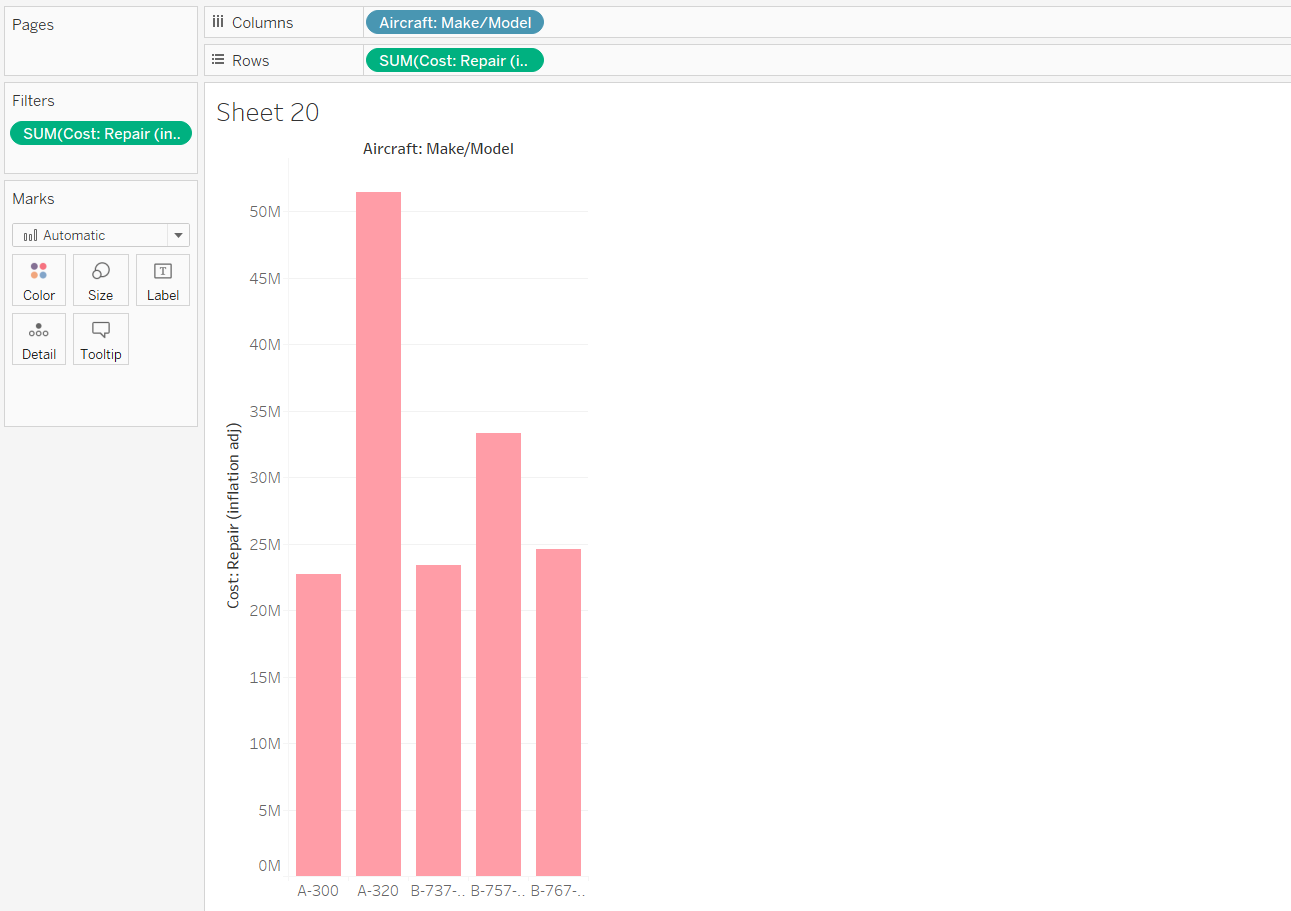
Sheet 10:



This sheet visualizes the Origin State,the size of the wildlife and the height above the ground. This will help:

1. Different airlines to keep track of the height above the ground(that is the height at which it travels)
2. End Users are-Engineers,geographers
3. Pilots

Sheet 11:

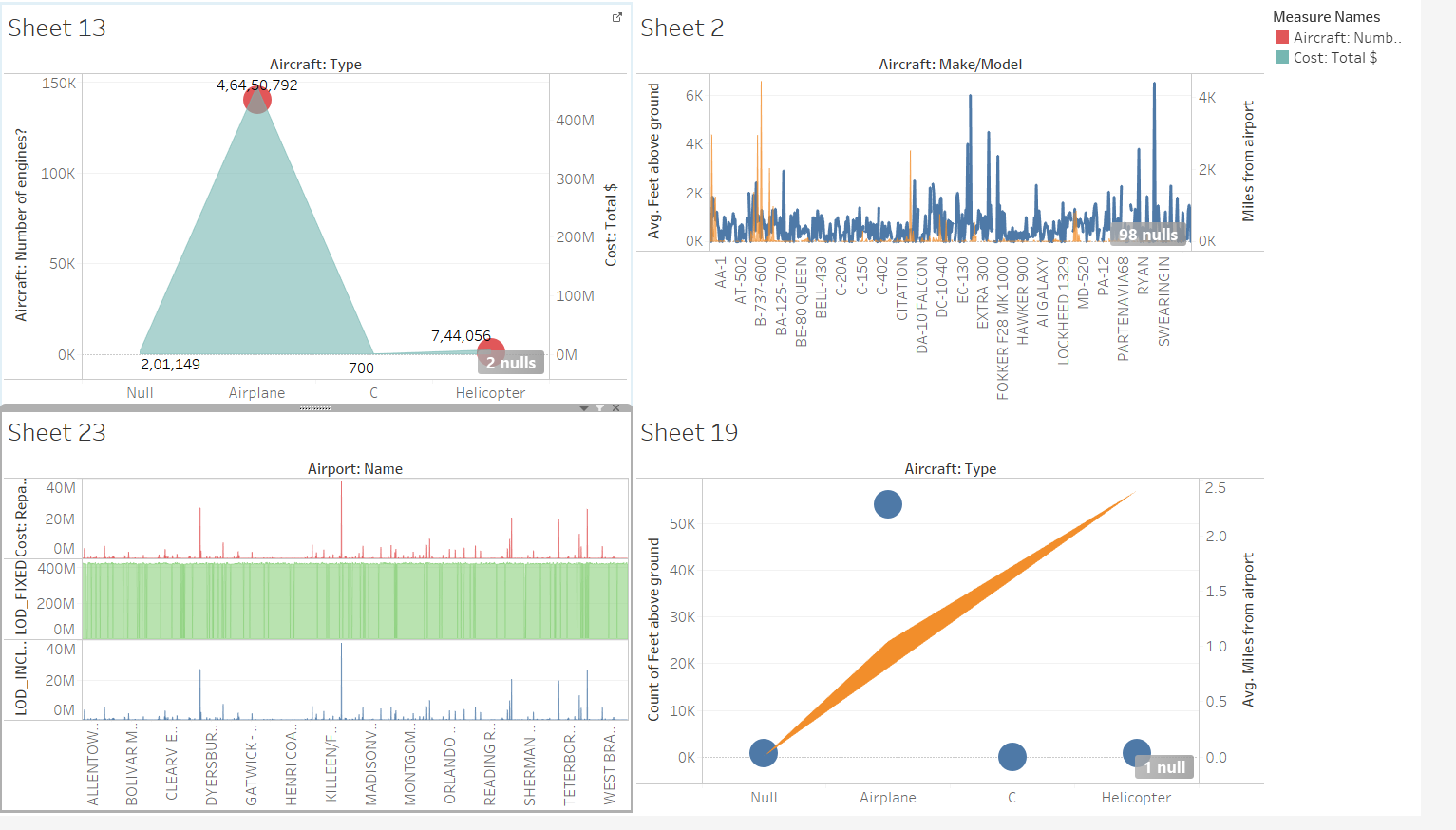


This sheet visualizes the model of each aircraft and their repair cost.This will help:

1. Different airlines to keep track of the cost of each model of aircraft.
2. End Users are- designers who make different models of aircrafts
3. Pilots

**Dashboards:**



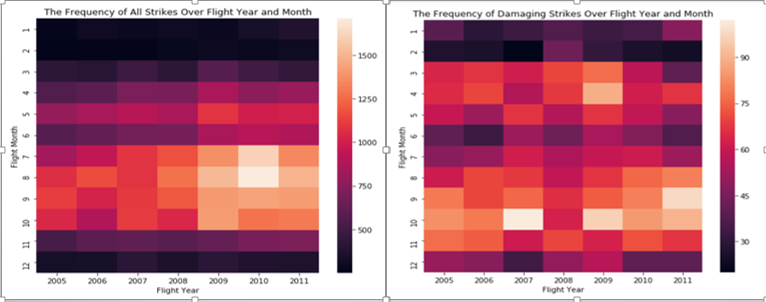


Inferential statistics

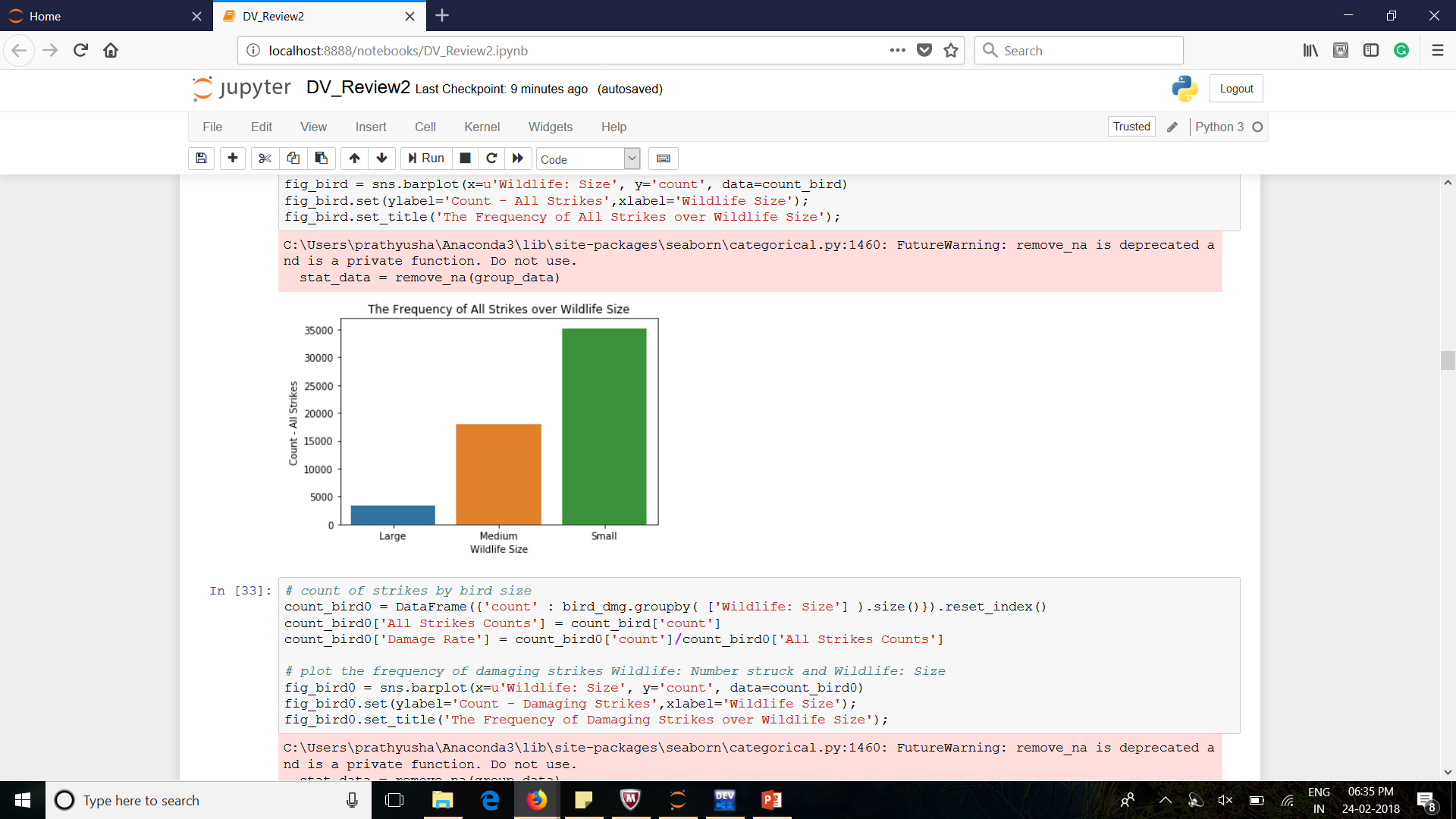
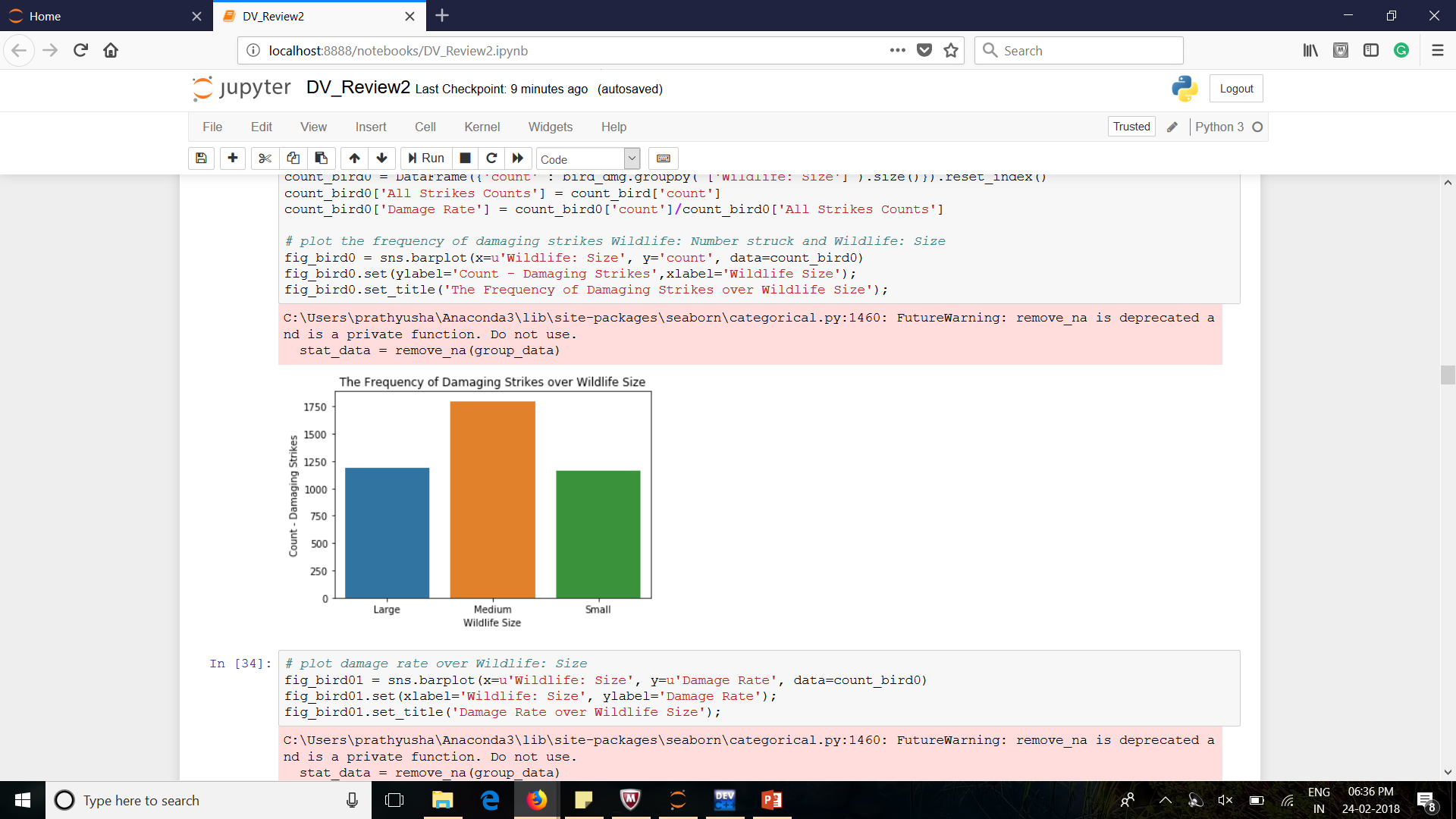
Damaging/non-damaging bird strikes will be classified using five models including

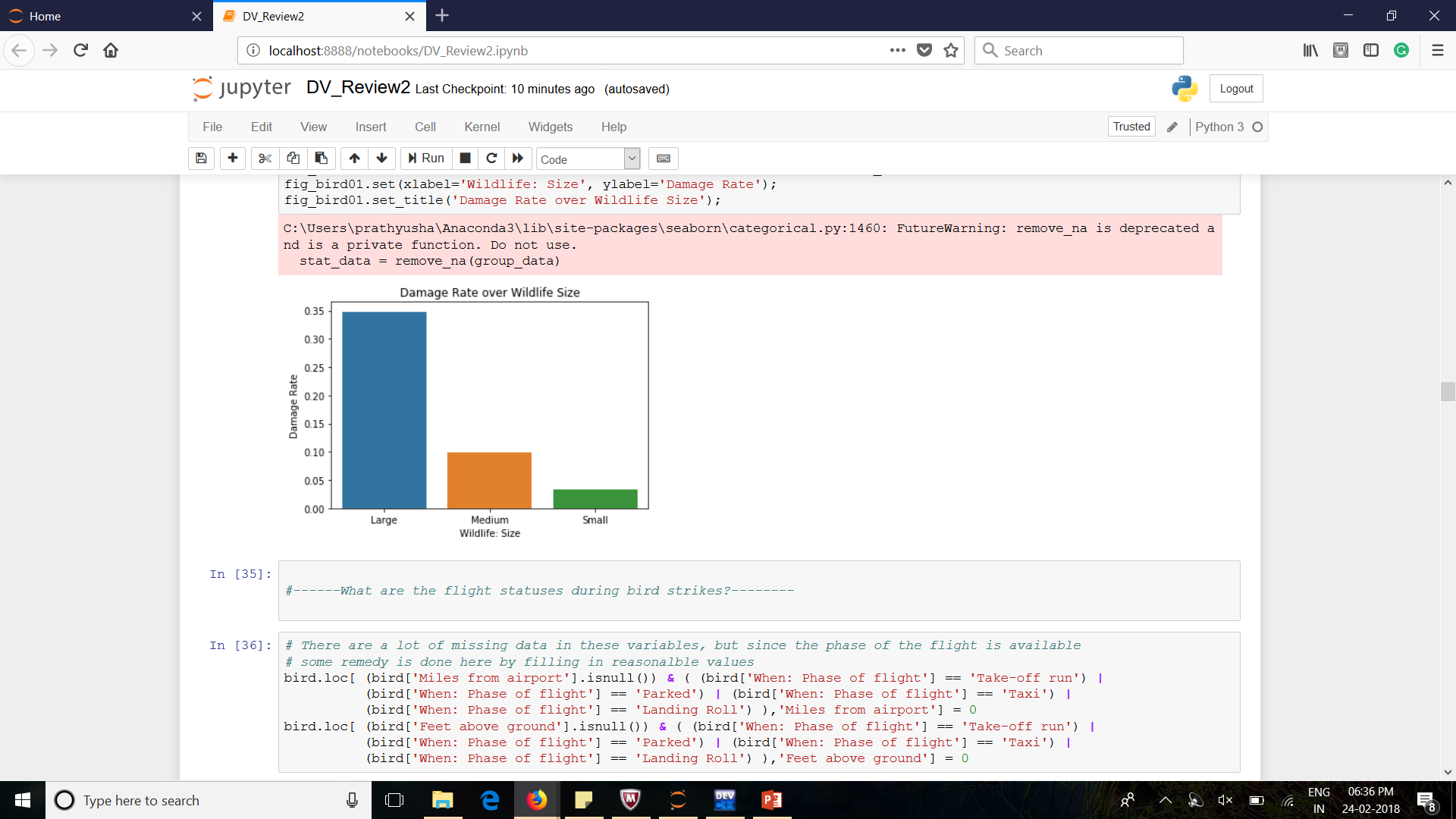
* + Logistic Regression
  + Support Vector Machines
  + Random Forests
  + K-Nearest Neighbours
  + Gaussian Naive Bayes.

Training accuracy, testing accuracy, and cross-validation accuracy will be used to check the fit of the model.



The two heat maps suggest that bird strikes happen year-round, mostly between *July* and *October*, increasing in frequency from year 2000 to year 2011. Damaging strikes happen between March and May, as well as between July and October, with a relatively stable trend over the years, suggesting that countermeasures take effect over the years.



Small birds are the most frequently involved in bird strikes. But medium and large birds cause damages more often.

**K Nearest Neighbours**

* Testing Accuracy: 0.631
* Training Accuracy: 0.882
* Cross-Validation Accuracy: 0.794
* ROC AUC on train set: 0.882
* ROC AUC on validation set: 0.687

**Gaussian Naïve Bayes**

* Training Accuracy: 0.668
* Testing Accuracy: 0.766
* Cross-Validation Accuracy: 0.678
* ROC AUC on train set: 0.743
* ROC AUC on validation set: 0.752

**Logistic Regression**

* Training Accuracy: 0.692
* Testing Accuracy: 0.781
* Cross-Validation Accuracy: 0.723
* ROC AUC on train set: 0.754.
* ROC AUC on validation set: 0.777

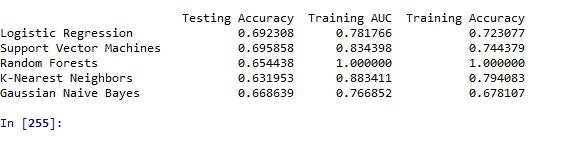
**Support Vector Machines**

* Training Accuracy: 0.695
* Testing Accuracy: 0.834
* Cross-Validation Accuracy: 0.744
* ROC AUC on train set: 0.802
* ROC AUC on validation set: 0.770

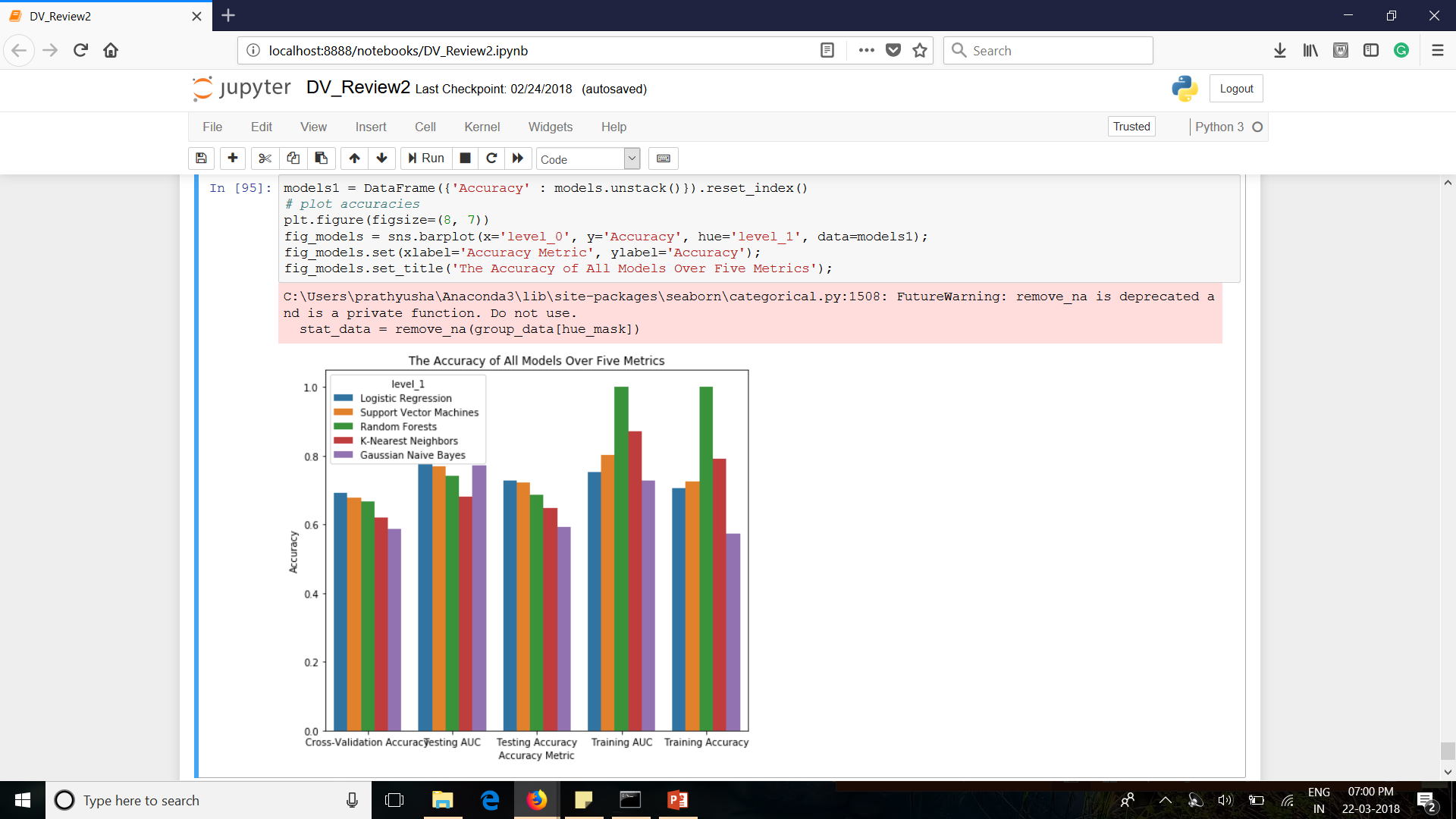
**Random Forests**

* Training Accuracy: 0.654
* Testing Accuracy: 1.000
* Cross-Validation Accuracy: 1.000
* ROC AUC on train set: 1.000
* ROC AUC on validation set: 0.748

**Model summary:**



**Accuracy of All Models Over Five Metrics**



**Conclusion:**

1. This project enabled us to identify various aspects of the dataset and act quickly to resolve any issues.
2. No. of airlines, Time of take-off, better customer services, location estimation, speed and fields like there can be predicted beforehand by visualizing the data.
3. In the end, we were able to see data from every aspect and apply various operations on it.
4. Based on the cross-validation and testing accuracies, the two most important metrics in model prediction among the three considered here, the ***Logistic Regression***model yields the best performance.
5. It is noteworthy that the ***Random Forests model*** performs slightly worse than the Logistic Regression on the metrics of cross-validation and testing accuracies. However, the Random Forests model has the highest training accuracy among all models. This suggests a potential model ***overfit***.
6. In summary, the **Logistic Regression Model** outperforms other models and can be chosen as a model for prediction and warning system for pilots and operators.