**PROJECT PROPOSAL**

**Topic:- Airline Data Analytics**

**Team members:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **User Name** | **First Name** | **Last Name** |  |
| sshreedharkulkarni | Sahana | Shreedhar Kulkarni | Voice of the team |
| ndarji | Nihar Anant | Darji |  |
| akong1 | Anna | Kong |  |
| bsoopal | Bhoj Rani | Soopal |  |

**Description:-**

We plan on working with airline datasets from kaggle website. We want to analyze data regarding domestic flights to predict flight cancellations in the US. We will analyze the cancellation including delay time, number of passengers, airline carrier, scheduled and actual departure and arrival times, and reason of delay [1]. Using these values we plan to future calculate delays caused by the airport, get the count of the flights delayed with respect to the total number of flights etc.

We also aim to calculate the average airline delay time, which airline has the most number of delays, at what time of the day do we see more number of delays in flights. We also want to see if there are any seasonal delays i.e if delays are mostly during the holiday season. We will use Spark’s MLlib to create powerful machine learning models alongside the DataFrame syntax and spark. We will set up on Amazon Web Services EC2 for big data analysis.

**References:-**

**[1]** Bureau of Transportation Statistics. (2016). Airline On-Time Performance and Causes of Flight Delays. Retrieved from <https://catalog.data.gov/dataset/airline-on-time-performance-and-causes-of-flight-delays-on-time-data>

**[2]** Deshpande, V., & Arikan, M. (2011). The Impact of Airline Flight Schedules on Flight Delays. *Manufacturing & Service Operations Management, 14*, 423-440.