**Work Plan**

**Data set description**

We are working on the airplane crash dataset for this project. This data set contains data about air crashes from 1908 to 2009. We have data regarding 5269 air crashes that have happened all over the world in different domains like commercial, military etc. Some of the attributes that our data set contains are the date the air crash took place, the time at which the crash took place, the route of the plane, the number of fatalities, the summary of the incident etc. We plan to make use of multiple attributes in order to analyze the available data and deduce some conclusions which can be used by our targeted audience.

**Research questions**

* Do more air crash accidents happen at a particular time of the day as compared to the others? For example, if more accidents are happening early in the morning or late in the evening (when it is not fully bright nor totally dark), steps can be taken to target specialized windows for pilots or lights for the airplane that can help counter this issue.
* Is there a region which is highly prone to air crashes? For example, Bermuda triangle is a region infamous for crashes and accidents. If it is established that there is a region where or around a huge number of accidents have occurred, either due the prevalent extreme weather conditions in the region or any other reason, the paths of the airplanes can be changed to avoid passing through those regions.
* Are human errors more responsible for air crashes than climatic or technical failures?  For example, intentional crash by a pilot, missile hit from ground etc. Have cases like this happened in the past? From these questions we can find out if things like pilot psychology need more attention.
* Is there a specific manufacturer of airplanes that is prevalent in many of the air crashes that have happened? If yes, all the airlines making use of airplanes of that manufacturer can be advised to exercise proper caution while buying more parts from that manufacturer.

**Target audience**

* Our analysis can have a wide variety of target audience ranging from the airplane travelers to air crash investigation agencies like the National Transportation Safety Board (NTSB). Air-crash investigation agencies would want to make use of our analysis to get a better idea about the reasons of air crashes which could range from technical failures to bad weather, to human intervention from the ground to intentional crash by the pilot. Based on the presence of patterns in the reasons of these crashes/accidents, appropriate strategies could be chalked out by such agencies to prevent such crashes in the future.
* The travelers can use this data to make sure which airlines they want avoid. If an airline or a specific airplane by a manufacturer is notorious for many accidents/crashes in a period of time, specifically for technical failure reasons, the travelers would want to avoid such an airline.
* The airline companies would want to make use of the results of our analysis to know which airplane manufacturer would they want to avoid. The airplane manufacturing companies like Boeing or Airbus would want to use our analysis to get the same information, but to make sure where they need to concentrate on, developing new airplanes with advanced technologies or improving on the existing airplanes.

**Workflow and Effort allocation**

**Data cleaning and Data Transformation**

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset. We found that there is a lot of data for various crashes that are missing in our dataset. In order to make sure that we have a reliable statistical analysis, we need to ensure that our data is complete. Hence, most of our tasks in the data cleaning phase will be focused on removing unwanted/incorrect data. For instance, there are missing values present in one of our attributes i.e. Route the flight took. This oddity might hamper our analysis if we want to analyze which route is more prone to air crashes, as we don’t have complete information for this. One way of editing this oddity can be searching for some more data sets related to air crashes and then filling out these fields using data from other sources. Another important issue related to data cleaning we need to keep in mind is removing incorrect data. For instance, there might be a few places where in the data entered might be incorrect. For instance, there might be a wrong entry of a text in the “time of crash” field, which is completely irrelevant.

Data transformation is the process of converting data from one format to another format. While we are interested in keeping the dataset in CSV file format itself, we are interested in adding new columns within this file for our reference, so that we can analyze the data more effectively. For instance, the reasons for the air crashes mentioned has nominal data, which needs to be categorized (as human errors, technical issues, weather and so on), since we are interested in analyzing the major causes. We can probably add another column for this, which has ordinal values for various reasons which are responsible for the crashes.

Effort Allocation: Since our dataset contains over 5000 rows, it will be an extensive task to find out the irregularities pointed above, and rectify them. We plan on dividing the dataset into three parts (one for each member of the group), and clean/transform the data accordingly. We intend to find out more on if we can automate this (probably using R), or will we have to do the same manually This will not only ensure that each of us understands the dataset extensively, but also ensures that we work on cleaning in parallel.

**Data Mining:**

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. It is during this phase that we will be looking out for information to answer our research questions. Since we will be mostly using R to extract information about patterns in our data, our focus will majorly be on generating R scripts and R plots for this purpose.

Effort Allocation: Since we have three research questions, we have decided that we will be taking one research question each, and analyze the above mentioned points for our respective research questions. We are yet to decide who takes which question.

**Presentation:**

The presentation phase mainly consists of integrating our various finings regarding the research questions. We intend to create a PPT with our findings.

Effort allocation: We plan on doing this together.

**Timeline and Gantt chart**

The below table describes our basic tasks and the approximate timeline for each task. Since our data set conains data only till 2009 we plan on looking for alternate data sets which might contain latest air crash incident details. This task will be completed under the task name ‘search for alternate data sets’.

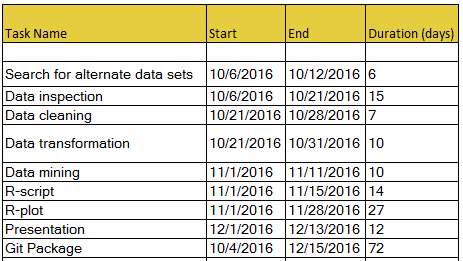
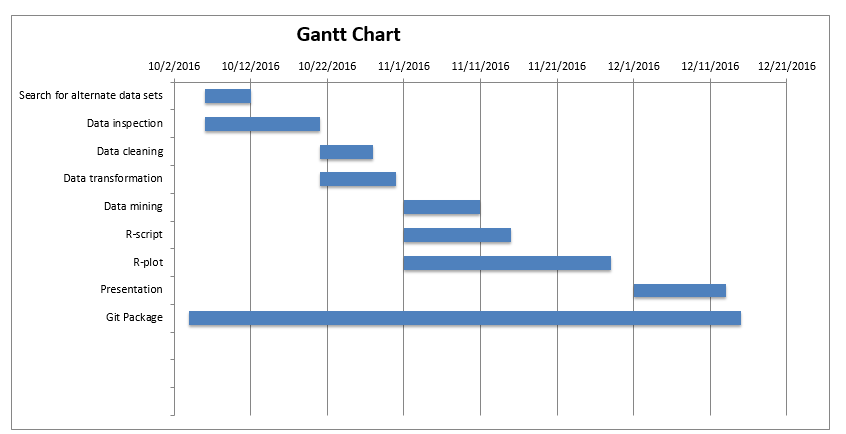
Table 1: Tasks

Figure 1: Gantt Chart

Word count=1177