Data DeCal Project Proposal Guidelines Due: March 4

Your proposal itself should be structured and **address all the following guidelines**. Also make sure you **submit the Dataset(s)** you wish to use with your proposal\*\*

1. Name(s) of who will be working on this project.

Enci Fang

1. State the problem/question you would like to answer through data science.

Which airlines are relatively safe?

Do we have to avoid taking the airlines that have had crashes before?

Which aspect do the passengers attach most importance to when measuring the airline?

Do the passengers’ positive airline reviews suggest the high-level airline safety?

1. Why is this problem important? What population is affected?

The main topic I want to choose on the project is the airline safety.

Recall the [downing of Malaysia Airlines Flight 17](http://www.nytimes.com/interactive/2014/07/17/world/europe/maps-of-the-crash-of-malaysian-airlines-flight-mh17.html?_r=0) in Ukraine, following the unknown disappearance of its Flight 370 in March 2014. Although such incidents don’t appear to be related in most cases, the high-profile crashes can robustly [shift passengers' demand away](http://www.finance.pamplin.vt.edu/faculty/vs/pdfs/JLE1998.pdf) from the airlines involved in the disasters. The fact offers me the insight into whether we should really reject to fly the airlines that have had crashes in the past and whether people who take such airlines will make bad comments.

The data-sets I collect is :

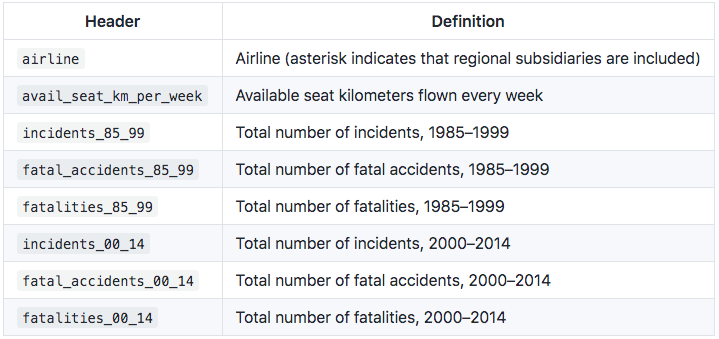
a. airline-safety.csv which includes the number of accident numbers from 1985-2014.

b. airline-review.csv which includes the ratings by random passengers

There are 56 airlines that were in the [global top 100](http://www.atn.aero/analysis.pl?id=1318) as of December 2012 and which have operated continuously since Jan. 1, 1985. As for passengers, it is useful to know which airlines are quite safer. If we find some relationships between the passengers’ review and airline safety, we may depend on the reviews to consider the choices of airlines. As for airlines, they can know which aspect the passengers pay attention to most during the flight. They can thus improve the resources concerned to make further development.

1. State the data/variables that you would need to collect to adequately address your problem (at least three).
   1. What you state should correspond to the variables in your Dataset
   2. In addition, explain **how** you will use the data to address your problem. Be as specific as possible!

**Q1:Which airlines are relatively safe?**



*(the description of data-set a)*

First, I need to sum the three kinds of variables, i.e, incidents, fatal\_accidents and fatalities, from two time periods to the whole period. Then calculate the incident rate, fatal\_accident rate, and fatalities rate. Intuitively, the lower incident rate, the much safer. Some airlines will encounter a few fatal accidents but cause great fatalities while some will get into a relatively large number of accidents but cause few deaths. So I plan to put these two categories together to get a weighted fatality rate. Also, the airlines with high avail\_seat\_km\_per\_week is likely to be more stable and safer than others.

Thus, for adjusted fatality rate and incident rate, subtract an airline’s crash rate from the average for all airlines. This gives safer airlines positive scores and less safe airlines negative scores.

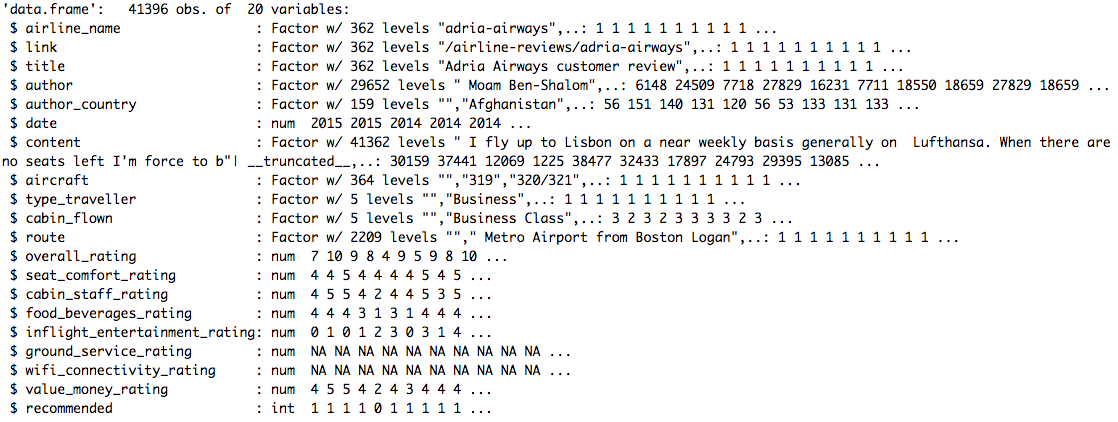
Multiply the result by the square root of avail\_seat\_km\_per\_week. This gives more credit to an airline that has achieved a strong safety record over a larger sample of flights.

[Standardize](http://en.wikipedia.org/wiki/Standard_score) the score in each category to calculate how many standard deviations an airline is above or below the mean. Then average the scores from the two categories together. This is the safety score. And finally take the airlines scored bigger than 0 as safer ones.

**Q2:Do we have to avoid taking the airlines that have had crashes before?**

Returning back to the original data-set.a, it has been already separated into two time periods. I can construct a linear model to see whether there is any correlation in the rate of fatalities from one period to the next. If the correlation is subtle, we can theoretically say that there is no need for us to never fly the airlines that have had crashed before.

**Q3:Which aspect do the passengers attach most importance to when measuring the airline?**

****

*(the structure of data-set b)*

We can roughly see that the passengers rates the airlines in these variables, seat\_comfort\_rating, cabin\_staff\_rating, inflight\_entertainment\_rating, food\_beverage\_rating, value\_money\_rating, excluding the missing data in ground\_service\_rating and wifi\_connectivity\_rating. I can use linear model and use F.test to compare the models to decide which one contributes most in overall\_rating by comparing the slope coefficient.

In addition, I can compare passengers flying the same airlines in different cabin\_flown to see if there is any difference for the overall\_rating. In other words, will the airline quality differ in the cabin standard? Here, I can choose one or two airlines to do the permutation test or t-test to compare the cabin-pair-wise mean difference in rating.

**Q4:Do the passengers’ positive/negative airline reviews suggest the high-level/low-level airline safety?**

Here, I need to combine the two data-sets. I will use the average rate of ‘recommended’ for each airline. Then make a plot of the ratings and safety score,or using linear regression to see whether there is a positive relationship between them.

1. What data science skills/topics will you use to complete this project (e.g. linear modeling, grouping, etc.)? Your answer here should correspond well to your answer in 3b.

Linear modeling. Permutation test. T-test. F-test. Grouping.

1. What programming language(s) will you use to complete this project?

R.

1. State the assumptions you have going into this project. (What is your null hypothesis, what do you believe about the topic, what do you predict about the outcome, etc.)

**Null:**

There is no correlation in the rate of fatalities from one period to the next.(Q2)

There is no difference in rating the same airlines with different cabins.(Q3)

**Assumption:**

All passengers have no knowledge about the history of airline crashes. And they make comments only based on the specific services they obtain rationally.

**Predict:**

The cabin passengers take may differ in rating the same airlines.

We probably don’t need to avoid taking the airlines that have had crashes before, since generally the risk of being involved in a crash is very low.

The passengers’ positive/negative airline reviews will to some extent suggest the high-level/low-level airline safety.

1. Set **at least 3** intermediate goals to meet over the course of the project (at least one goal for each semester checkpoint).

Data cleaning and combining the data sets.

Completing Q1 and Q2.

Completing Q3 and Q4.

1. Discuss some possible limitations of your project this semester? What do you think will challenge you? Note that these limitations should not interfere with your overall study on the dataset.

The assumption may be violated because it is impossible for each passenger to have no knowledge about the airlines. People knowing the previous crashes may subjectively make extremely bad comments on the airlines no matter how good the service it provides. In that way, we cannot rely on the passengers’ review to find the relationship between the reviews and the safety of airlines since there exists the internal relationship between them. But here there is no variables concerning safety standard. I guess the passengers are interviewed based on the variables in the dataset, especially there are the text comments they make in the ‘content’ variable, matching the high/low ratings in certain given aspects.

I think the most challenging part is that when combing two data-sets, since one airline may have different name formats in two data-sets, I need to do some regular expressions to match those airlines. Actually I am not sure whether it works. I sincerely hope that I will get some help from instructors at this point.

**Grading Rubric**

|  |  |
| --- | --- |
| The proposed problem can be studied by the provided dataset. | 7 Points |
| Discussed the problem’s implications and why it is important to the population it affects. | 2 Points |
| Dataset has a valid number of variables to work with. | 2 Points |
| Valid data science tool to use for study. | 2 Points |
| Programming language proposed. | 1 Point |
| Stated assumptions. | 2 Points |
| Set goals for each project checkpoint. | 2 Points |
| Stated reasonable limitations that don’t impede with the overall dataset study. | 2 Points |

20 Points

After the submission due date, you’ll receive feedback and a score based on this rubric. If you don’t get full points, but fix your proposal based on our feedback, you’ll receive the points back.

***Any significant proposal changes made after April 15 (2nd semester checkpoint) are subject to a permanent loss of points on your proposal score. This is to ensure that each person’s project is appropriate and ready to present by final presentations.***