**Project Proposal Form**

**Title of your project proposal**

Why is my flight delayed?

**Team Name**

BBQ

**Team Member 1's Github username**

davidmidd

**Team Member 2's Github username**

egranet

**Team Member 3's Github username**

foo-bar-baz-qux

**Team Member 4's Github username**

LeonhardFS

**Background and Motivation**

*Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.*

Our project is motivated by the following facts:

* Flight delays are an issue for many business and official travellers. Over the years, some airports such as Washington Dulles, DC have earned a reputation for being unreliable. Based on the data, we want to explore whether these reputations are deserved.
* There is extensive official data on the topic of flight delays. More than 8 million commercial flights are operated every year in the US. For each of these flight, information such as date, weather conditions, airline carrier, departure and arrival airports are freely available.
* Numerous charts can be plotted. For example, what is the link between the political color of a given state and its performance against the national average in terms of flight delays? How about the relationship between airport location and city center (as given by Google Maps)?

**Project Objectives**

*What are the scientific and inferential goals for this project? What would you like to learn and accomplish? List the benefits.*

We are pursuing two objectives:

* Predictive inference. Our aim is to build a delay estimator. Inputs are the airline carrier, the departure airport, the destination airport, the time and date of the flight, weather, and other variables deemed relevant by a statistical analysis. Output is the expected delay. Thanks to the high volume of data, we might even achieve a probability density function for the delay. This would require statistical tools such as non-parametric kernel estimation, linear and nonlinear regressions. As for computations, the large amount of data would require to use cloud services such as Amazon Web Services.
* Explicative inference. Understand what causes delays. Do some airports systematically generate delays independently of the company? What is the influence of seasonality? Is there a state factor? When a flight is delayed, what are the repercussions on following flights? Some plots seem challenging. We still have to figure out how to color a map of the US without going state by state on Paint...

**What Data?**

*From where and how are you collecting your data? Is the data publicly available? How big is it?*

Since 1987 the U.S. Department of Transportation (US DOT) collects a huge amount of domestic flight data. This dataset, containing many different features (e.g. timestamp, origin, destination, airline, delays, etc.), is publicly available on their website <http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236>. The tool on their website allows for downloading a customized dataset based on selected features and a specific time period. Each month there are more than 400,000 recorded flights, so we have to deal with very big datasets (or just focus on a subset).

Furthermore, we might be able to connect this dataset with external data sources that affect flight delays (e.g. weather data, strikes, etc.).

**Must-Have Features**

*These are features or calculations without which you would consider your project to be a failure.*

Required features and calculations are:

* a classifier that tells us whether a flight will be delayed by more than 10 minutes.
* various correlations plots between variables such as airport and carrier.
* various maps of expected delays by seasonality and state

**Optional Features**

*Those features or calculations which you consider would be nice to have, but not critical.*

* a probability density function estimator. Given a bunch of input parameters determined by our statistical analysis, the predictor yields the pdf of the delay.
* outlandish correlations such as political color and average delay.
* a web-based online tool which allows users to retrieve the statistics in an intelligible way

**Design Overview**

*List the statistical and computational methods you plan to use. Are you planning to use AWS?*

We still have to figure out which part of the data is necessary to produce high quality results. As outlined above, the dataset can be very large (e.g. >4 million entries for the last 10 years). If we use the complete dataset, we will certainly consider to do some part of the analysis on scalable systems such as Amazon AWS.

We intend to explore a wide range of different classification and regression techniques to see which algorithm performs best, and what can give us the most interpretability for drawing inferences (e.g. logistic regression, SVMs, and random forests).

**Verification**

*How will you verify your project's results? In other words, how do you know that your project does well?*

Since the flight delay dataset is quite popular within the data science community, there are already some analyses on the web that could provide us with opportunities to benchmark our results. Furthermore, we will discuss our approach with the TF and will report on important updates regarding our project.

**Visualization & Presentation**

*How will you visualize and communicate your results in your video and website?*

For the exploratory analysis, we plan to use a combination of matplotlib graphs and integrated interactive visualizations based on Tableau Public that we will try to embed into our webiste. The website will certainly be structured into several tabs (e.g. a differentiation between descriptive and predictive analysis) to present our work regarding each of our main goals.

**Schedule / timeline**

*Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.*

Week 12 (11/16-11/22): Finished project plan and defined key questions, all data sources identified.

Week 13 (11/23-11/29): Data acquisition and cleaning completed. First exploratory analysis.

Week 14 (11/30-12/6) : Visualization of exploratory analysis results finished. Start of predictive modelling (experimentation of different models).

Week 15 (12/7-12/13): Finish predictive model. Website setup and video recorded.

Week 16 (12/14-12/20): Final presentation

**Team Member Contributions**

*List the contributions each team member will make.*

Estienne: Data cleaning, feature extraction, exploratory analysis, refining computational and statistical approach, implementing a predictive model

Leonhard: Exploratory analysis, feature extraction, visualization, refining computational and statistical approach, implementing a predictive model

Victor: Refining key questions, visualization, implementing a predictive model, working on presentation and write up, helping with website

David : Visualization, implementing a predictive model, creating and refining website, video recording