* **Tell us about your project**
* Background and Motivation \*
* Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

(revised)

The process of deciding on a project took much more time and effort than expected. We found brainstorming dozens and dozens of interesting ideas to not take too long; however, many of these ideas were not viable project ideas, due to one or more specific reasons: they focused more on mathematical modeling than data science, the necessary data was not publicly accessible, the project required too much background knowledge to be mastered before actually beginning the data collection/analysis process. Some members of our group have interests in transportation, and one idea we had was about trying to find trends in flight ticket prices in the months, weeks, and days leading up to the actual flight. Upon studying what attributes of a flight led it to have specific trends in price, we could then predict, given attributes of a new flight, when the prices would likely be the lowest (and thus when it would be most advantageous to purchase a ticket).

Unfortunately, we were unable to find publicly accessible data for historical prices of flight tickets; data sources found online only included the final ticket prices of past flights or today's prices for upcoming flights. During our search, however, we came across the flight-related databases from the US Department of Transportation's Bureau of Transportation Statistics, which track various attributes of all commercial domestic flights from recent years. Among these attributes was information about scheduled and actual departure and arrival times for flights, as well as reasons for delays. This inspired us to pursue a project related to flight times and delays.

Around the same time we also came across the video presentation for a CS 109 project from last year, which coincidentally used the same dataset from the Bureau of Transportation Statistics. In their screencast, they discussed how they studied the relationship between flight delays and weather patterns. After scraping weather data for each day/time of day in different parts of the country, they used binary indicator variables for each flight to signal whether it was delayed or not, focusing on the difference between typical flights and flights that were significantly delayed as a result of weather. They noted that after much optimization, their most useful and promising model, which used random forest techniques, still had little success predicting which flights would be delayed, with an approximately 10% recall rate. It made some intuitive sense to us that despite the team's expertise and efforts, working with weather yielded suboptimal findings, since most prolonged delays probably only occur under severe weather conditions, and prediction of that relates more to meteorology and not flights. Plus, in these cases, all flights would consistently be delayed or cancelled; this would greatly restrict the variability between different flights, which we thought would be most interesting to study.

When discussing one of their initial approaches with linear regression, the group noted that little insight could be derived about long-term delays due to weather, because the data was polluted by the much higher frequency of short-term delays that had no correlation with weather. We decided to consider these delays in our project, looking even at how the variability between flights that are 10 minutes early, 10 minutes late, etc. can be explained by attributes of the flights. What explains the variability, instead of weather, which seems not to correlate at all? We may consider longer delays as well, however we expect that to be correlated to more external, unpredictable factors such as airport security breaches or severe weather patterns.

Note: We did not review any part of this other group's final project other than the screencast, deciding it would be preferable to come up with all analysis techniques ourselves. We figured that although our project is different, the related nature of their project would suggest specific approaches, and that might hinder our own exploration and learning, which is, at the core, the purpose of this final project.

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

(revised)

We intend to study what factors related to a flight might influence how long it takes and how that relates to the predicted departure and arrival times. These thousands of causes are mostly going to be impossible to enumerate, quantify, or measure reasonably (ex. how bigger overhead carry-on bin spaces make it take less time for passengers to get settled). However, many of the more important causes, as well as many attributes of flights that are easily measurable, are recorded in the government database.

By analyzing how the various flight attributes correlate, we can see which ones are most related to how long flights take compared to what the airline predicted. Based on how these results turn out, there are many possibilities for how to proceed, including using regression to predict the delays on a new set of flights, using the most important attributes in machine learning classifiers and then predicting on a test set, considering practical implications for airlines, and incorporating information from other databases.

Although the implications of our results cannot be known in advance, this project is worth completing, even though at first, it may seem that small differences in flight time or very short delays may not be very interesting. Firstly, we hypothesize that elementary statistical methods will be able to shed more light on flight variability when studying more consistent factors (unlike weather, security breaches, etc.) Second, less attention has been paid to studying these more fundamental attributes of flights, so our insights may be more unique or valuable. And finally, these findings could actually be of consequence to the airline industry. Even if the magnitude of time differences for individual flights is small, there are over 30,000 domestic flights daily, so effects can add up. It would be in the airline company's interest to study how they can better predict when flights will arrive and depart and how they can increase speed and/or efficiency of their flight schedules. Perhaps they make regular overestimations or underestimations of flight time due to specific factors that can be only discovered through statistical analysis. Such knowledge might influence them to fly through certain airports more, revise predictions to be able to make delays less common, decrease time that airplanes are sitting idle on runways, or make any variety of other changes. The airline carriers are our primary audience, but ultimately the scale of how busy and complex our commercial aviation industry is motivates how possible insights could be in the nation's and everyday passengers’ interests, and regardless of what we discover, makes our project interesting to pursue.

* Must-Have Features \*
* These are features or calculations without which you would consider your project to be a failure.

(revised)

We should consider at least five different attributes of flights, using data from the Bureau of Transportation Statistics data tables, and study how much the arrival times of the flights differ from what was scheduled. We may discover that we need to simplify the data set along the way, such as considering a few specific routes from one city to another instead of all flights. We should report some form of numerical value about the relationship between each attribute that we study and the target metric, most likely through correlation from a regression model. We should also consider at least one, but probably multiple, analysis of more than one factor at once, looking for clustering. We should make qualitative interpretations/explanations in addition to having the numerical values as well.

From this, we should decide on one or more attributes that seem to correlate either surprisingly well or at least better than the other attributes to complete at least one, but likely multiple, suitable machine learning techniques on. Based on what category or value (for that attribute) a flight belongs in, we want to make a prediction as to the expected difference between scheduled and actual arrival time. We should compute metrics on how successful our machine learning approaches are, much like in Problem Set 5, and then again make qualitative observations about this performance.

There will of course by deviations as we complete the project, but the main idea is that we want to generate regressions, identify factors with higher correlation, apply machine learning using those factors, and interpret performance of our methods in a broader context. More technical specifics are discussed in the Design Overview section.

* (Evan) Optional Features \* (brainstorm things to have as second-layer analysis questions)
* Those features or calculations which you consider would be nice to have, but not critical.

After completing our analysis of delays in US domestic flights, we could possibly turn out attention to other countries whose data is readily available. By comparing different countries, we might be able to draw conclusions about how various government/airport policy or cultural efficiency plays in predicting flight delays. For example, we might see that flights in Japan or Dubai are especially on time because of the Japanese people’s inherent punctuality or the strict timetable of Dubai’s airport.

Other features that we may analyze include the cardinality of travel, flight duration, and taxiing time. Wind patterns along with the direction of the flight may affect the flight’s landing patterns. Flights with longer flight times have a higher chance of “making up” lost time in the air. Finally, time spend on the ground during taxiing can also be analyzed to better predict exactly when a flight will arrive at its gate.

Optional visualizations include a heat map of the United States showing where the most delays are occurring, some sort of interactivity on the website whereby people can look up / search for flights that have already landed and see our predictions. We might not be able to get a website with interactivity, but the iPython notebook will definitely be able to accept user input. This would require a real-time database like flightaware.com.

Consider data from other countries to compare? Shouldn’t be too hard to find.

Discuss implications for how airlines could make changes?

Something related to cardinal direction of travel and how that impacts flight times/delays. (Ex. why flying NY to LA and LA to NY take very different amounts of time.) Might need to look into wind patterns.

Considering alternative metrics in addition to difference from predicted flight arrival time. Perhaps worth considering include differences in total duration of flight compared to other flights with the same origin and destination airports, differences in departure time from expected, or differences in taxi in/out time compared to expected/average.

Adding cool visualizations related to the factors we find most important? Ex. idea with map of United States.

Adding interactive aspects to website? Being able to predict flights in real time by having user-friendly input. Or maybe not on website, but at least in iPython Notebook. Would require working with a real-time/recent flight database and making data from it compatible with our data.

How do delays spread throughout the transportation network? Can we predict whether a flight will be delayed based on OTHER flights in the same airport and their on-time status?

* What Data? \*
* From where and how are you collecting your data?

(revised)

Our primary data source is from one of the Bureau of Transportation Statistics' Transtats databases, which contains on-time arrival data for major carriers’ commercial domestic flights. Information in these data tables include departure and arrival delays, info about the flights such as flight number and carrier, as well as some data on causes for delays and cancelled/diverted flights. [<http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236>]. Upon selecting a time window and choosing which of the 109 attributes one wishes to include in the table, one can download a .xls or .csv file. It is possible we will use other data sources later on, especially in order to implement some of the optional features.

In particular, we will likely be interested in data that tracks very recent and upcoming flights. On such website is flightaware.com. They provide an API which we can use to check the scheduled arrival / departure times of any flight based on its call number. We expect that data from flightaware will be just as accurate as that from the BTS and that it will be updated much more quickly (in real time).

* Design Overview \*
* List the statistical and computational methods you plan to use.

Our exploratory analysis phase will involve examining the effects of the following input factors on a flight’s delay time:

* Flight time (hour/min) -- treated as continuous
* Day of the week -- categorical
* Date (of the year) -- treated as continuous
* Airline carrier (categorical)
* Origin airport/Destination airport -- We will both try controlling for this variable by looking at data with a single origin and destination airport pair. We will also try examining the effect of this variable by splitting into categories, such as airport size, location in the US, etc.
* Total distance of flight -- continuous
* Plane type/passenger capacity.

Flight delay time is assigned as some combination of the following five delay times: carrier delay, weather delay, National Air System delay, security delay, late aircraft delay. The sum of these times is the total delay time for an aircraft.

For the continuous inputs, we’ll start out by looking at scatter plots between the inputs and the delay times, looking for possible significant relationships by eye, and perform linear regression to detect correlations. We’ll also do the same to examine possible significant relationships between the continuous inputs, and we’ll see if treating the delay as a categorical variable (delay/no delay) leads to any significant results using logistic regression with the continuous input variables.

For the categorical inputs (day of the week, airline carrier, type of airport), we’ll look at scatter plots of delay time vs the categorical input to see if the delay times tend to cluster with respect to the categorical input, and also look at histograms to look at the frequencies of various bins of time delays with respect to the categorical inputs.

We’ll further look for clustering behavior by looking at a histogram of delay times and see if they tend to cluster together around certain times. We’ll also try using cluster analyses such as k-means clustering to examine whether times that cluster together in the histogram mentioned above also cluster together when examined in various combinations of dimensions of the data.

Based on these initial analysis, we’ll examine whether it would work to bin the delay times, which would enable us to explore how machine learning techniques perform in predicting delay outcomes. If we do find that delay times tend to cluster in some n-dimensional space of relevant inputs, KNN and random forest (or simple decision trees) may be effective in forming a model to predict delay time outcomes. If the initial analyses indicate that it would work to bin the delay times into just two bins, we could also explore using support vector machines.

(Evan) Verification \*

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

Our database with the BTS has a ton of data. Therefore, we can afford to split it up initially into training and testing sets. Given a specific flight from the testing set that we want to predict, we will take data appropriate data from our training set to train our machine and then output a prediction for that particular flight. We will do this for many flights in the testing set in order to come up with the overall aggregated accuracy of our model.

Additional verification (and more interestingly), we could use real-time flights in the air that have not yet landed. We can use our training set to predict their arrival time and then compare it with their actual arrival time in the future.

Additional verification could come from: using very recent flights obtained from other databases to evaluate our predictions. Then, even using real-time flights in the air to make our work applicable in a real situation.

We will try to predict when they will depart/land and test it against the actual results that we will receive one the takeoff or departure actually happens.

* (Evan) Visualization & Presentation \*
* How will you visualize and communicate your results?

We will be showing many visualizations and graphs in our project. First, we will be performing exploratory data analysis on features of the flights that have the most impact on the actual arrival time of the flight. We will show scatter plot graphs with linear regression to weed out the explanatory variables and remove irrelevant variables. We will also look at how data tends to cluster in order to group similar flights together in our analysis.

After we finalize which features are interesting, we will perform machine learning techniques to classify flights based on those features into delay or not delay. Then, depending on which technique we use, we could show a decision boundary (for just the 2 most important variables), or other types of visualizations suitable for the technique we use. As an optional feature, we hope that our website will have some kind of interactivity to it so that users can directly input the flight they would like to check. Such a visualization might give a percent probability that a particular flight will be delayed. We will also make visualizations that summarize how accurate our model is: perhaps some kind of scatter plot comparison with different colors representing correct or incorrect and the axes representing 2 features that we found were prominent. Other interesting possibilities are looking at all the flights coming into/out of a particular airport and making a heat map of the US based on which airports are experiencing the most delays.

--linear regression/scatter plot of the most powerful explanatory variable vs. irrelevant variable

--interesting clustering examples

--depending on what factors are relevant, other visualizations such as a map visualization

--prediction tool based on entering flight number

--via ipython notebook

--perhaps PPT/website view as well as scrolling in iPython notebook? To make it cleaner looking and easier to see all graphs/visualization and highlight most important conclusions

* Our website will hopefully include the following features:
* **Delay Estimator**: the user can input various information about a particular flight and we will calculate the expected amount of delay that flight will experience
* **Airport Checker:** given an airport, it will aggregate the current delays and predicted delays to give a rating
* Will Elaborate...We will also provide a cool visualization whereby users can see the historic delay of this particular flight, color coded by length of delay.

some of these in iPython notebook instead? Then maybe have in website in more jazzed up, interactive, visually appealing way

[be a bit conservative on this, and then say “here are some other good ideas we might implement” haha]

* 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.
* **By end of Thanksgiving 12/1:**
* Exploratory Steps (get help from TFs)--figure out what kind of relationships are salient/significant, which features to include/exclude, find clustering
* **Between 12/1 and 12/8**
* formalizing--trying out ML techniques and comparing them, linear regression model
* **12/8 -12/10 - 12/12**
* Finalizing ipython notebook, creating website and visualizations