**[Final Project Submission](https://blackboard.syracuse.edu/webapps/assignment/uploadAssignment?content_id=_5322499_1&course_id=_393757_1&group_id=&mode=view)**

**Deliverable**: One R source code file, with a filename that includes your last name, containing analyses and descriptive comments as specified below. The R code must run, from start to finish, without errors, using the dataset provided. One MS-Word file, containing a three- to five-page report to the project sponsor that documents your analytical results.

**Rules of Engagement**: This is an honor system assignment: You **may**consult with me, the textbook, and publications on the Internet at any time. You **may not**consult, collaborate, or seek assistance from any other human besides me. Your attribution statement, at the top of your R-code file, must reflect these constraints. You may not share your results or work in progress with any other human besides me. Note that your data file is unique to you: The results that other students in IST687 obtain will be different from yours. Consult the schedule at the end of the document for the recitation sessions at the end of the semester dedicated to debugging but you will be happier if you get an early start on the project. Reminder: if you don’t come to the recitation sessions and demonstrate your progress for each milestone, you may not seek additional help during office hours.

**Project Goal**: The goal of this term project is for you to use all of the skills you have developed in the labs/homeworks to make sense of a novel dataset, to perform some essential analyses on the dataset, and to explain/document what you have done. The dataset contains summaries of air travel within the U.S, one row per trip.

**Accessing Your Data File**: Your JSON file will be mailed to you no later than Thursday, October 24th. The JSON file should contain about 10,000 observations and about 32 columns/variables. Each row represents one customer’s airplane trip from an origin to a destination.

**Five Project Phases**

* *Phase 1: Mitigate Missing Data.*There are several columns in the dataset that **may** contain missing data. Write code that examines each column to see if it contains missing data. To mitigate missing data, use mean substitution for numeric variables. Use comments in your code to document how many missing data values you had to repair.
* *Phase 2: Summarize variables*. For each numeric variable, create a histogram. Add a comment that describes the shape of the histogram as symmetric, positively skewed (long right tail), or negatively skewed (long left tail). For each factor variable (e.g., Gender), use the table() command to summarize how many observations are in each category.
* *Phase 3: Predict Satisfaction from Other Variables.*Many columns contain data relating to the characteristics of each customer’s trip. Some of these columns are numeric and so will be easy to use with lm(). Other columns are string or factor data and may need to be converted to be usable with lm(). You can use as.factor() or as.numeric() to coerce character data for use with lm(), although lm() is pretty smart about coercing data when it needs to. Remember the guidelines for interpreting lm() results: Interpret the adjusted R-square first, then interpret any predictors whose “Estimates” are statistically significant (indicated by one or more “\*” at the end of the line). From the standpoint of communicating with the client, it can also be useful to run a second “trimmed” lm() model that only includes those predictors that were significant in the initial model.
* *Phase 4: Map Low Satisfaction Routes.*Subset your data to create a smaller data set containing only the trips where customers reported the lowest levels of satisfaction. The latitude and longitude of each origin and destination is shown in the data set. Use ggplot to place route curves onto an outline map of the U.S. states. The geom\_curve() geometry supports this kind of plotting. You could optionally try to represent an additional variable on your plot, for example by driving the color of the curved lines with a numeric variable.
* *Phase 5: Make Sense of Low Satisfaction Segments*. The client wants to know why customers become dissatisfied with their air travel. Use insights from Phase 3 and Phase 4 to explain why certain trips have low satisfaction. Conduct any appropriate follow-up analyses to provide evidence for your ideas. Make sure to document any additional code with appropriate comments.

Your three- or five-page report should provide the client (presumably the leaders of the airline) with an explanation of your results in language that is suitable for a manager to understand. Most managers do not know too much about statistics, so you probably should not quote terms like “R-squared” or “p-value” but rather describe your statistical results in plain language. Your report should contain at least one to two graphics, presumably from Phases 3, 4 or 5, that illuminate the message(s) you are trying to communicate to the client. **Your report should conclude with at least one substantive recommendation to the managers on how they can improve satisfaction.**