# STAT 19000 Project 12

## Topics: Fun with Pattern Matching

Motivation: We close the fall semester with a second project on pattern matching. Regular expressions are a useful way of quickly classifying data into groups. They also allow us to do exploratory data analysis in a way that would be hard to accomplish otherwise.

Context: In this project, we use regular expressions as a way to systematically classify data. We will come back to this topic in the spring semester, as we wrap regular expressions into other kinds of functions for text analysis.

Scope: It would be useful to print The Basic Regular Expressions in R Cheat Sheet from RStudio and to keep it handy during this project. Even though it says "R Cheat Sheet", the regular expressions found here will be useful in other languages and platforms as well.

https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf

How to format your solution: Use the `project12template.R` file from the `/class/datamine/data/examples` folder.

## Question 1: AirBnB reviews in Los Angeles

Consider the data from the Los Angeles AirBnB "listings.csv" file. (Remember that we first explored this data in Project 5.)

1a. Among these five words, which word appears most often as the first word of the "name" column: "Beautiful", "Charming", "Cozy", "Modern", or "Private"?

1b. Among these five words, which word appears most often as the last word of the "name" column: "Apartment", "Beach", "Hollywood", "Home", or "House"?

## Question 2: Classifying Tailnums

Consider the tail number column, from the flights in the 2019 data set. There are 2997647 entries in this column.

We use the phrase "alphanumeric" to describe a character that is a digit 0-9 or a letter A-Z.

How many 5-character tailnums are there, with the form: `N`, followed by 3 digits, followed by 1 alphanumeric character?

How many 6-character tailnums are there, with the form: `N`, followed by 3 digits, followed by 2 alphanumeric characters?

How many 5-character tailnums are there, with the form: 3 digits, followed by `NV`?

How many tailnums a there, with the form: `ALL`?

How many tailnums are blank?

[Hint: Make sure that the total number of tailnums agrees with the number of rows in the data frame. There are no overlaps among the 5 categories mentioned above.]

## Question 3: Analysis of Amazon music reviews

Answer a fun question (of your choice) about a musical artist who you like, using the file of more than 4.5 million Amazon music reviews in the file:

```
/class/datamine/data/amazon/music.txt
```

Hint: If you want to read in the file to R, you can use this code:

```
myDF <- read.delim("/class/datamine/data/amazon/music.txt", quote="", header=F)
```

## Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/AVwpVVv7 using the instructions found in the GitHub Classroom instructions folder on Blackboard.