title: “Capstone Project”

# Problem:

Insurance companies rely on policy revenue for their profits. When customers cancel policies prior to a year, there is lost revenue. It is difficult to prevent this loss of revenue because we do not know who will cancel ahead of time.

# Uses:

An insurance company could potentially use this information for two purposes:

1.Identify the leads that are likely to cancel and do not buy those leads. That money would be better spent of buying leads for customers who will be less likely to cancel.

2.After a policy is purchased, identify those who are most likely to cancel. Then have targeted interactions with them aimed at preventing cancellations e.g. payment reminders, phone calls to ensure satisfaction.

# Data:

Data is available from a small insurance company. The data being used contains the following variables that will be used to examine cancels:

\*Geography: State, Zip  
  
 \*Demographics: Gender, Age, Marital status, Occupation, Credit Category (poor, average, excellent)  
  
 \*Previous insurance: Insured vs. not insured, Carrier Name, How long until policy expires, Policy type (monthly, annual, etc.), BI limits, Deductible amount  
  
 \*Car: Make (Nissan, Buick), Annual mileage, Number of vehicles  
  
 \*Marketing: Marketing campaign (Leads are purchased from different marketing sources)  
  
 \*Purchase Information: Date of purchase, Insurance agent name, Same day vs. follow-up purchase. Purchase data would only be relevant for preventing cancellations post-purchase, not for lead purchase guidance.

The data can be examined by lag (time from purchase to cancel) and reason for cancellation (non-payment, competition, etc.)

# Approach:

1. Data Wrangling: Make sure data is organized in the best way possible:

* \*Combine data sources: Data will be obtained from three sources: Lead data file, Purchase data file, and cancel data file
* \*Examine missing data. If some leads have too much missing data, eliminate them from the data set.
* \*Turn lag time into a categorical variable (0-90 days, 91-180 days, 181-364 days)
* \*Date Transformation: Create categories for examining date: Month, Day of week
* \*Any other issues that need to be addressed before analysis?

1. Exploratory Analysis:

* \*Which variables correlate with cancels?
* \*Graphs to show variables of interest

1. Predictive Analysis:

* \*Anaysis to show which variables predict cancels
* \*Analysis to show which variables predict cancels by lag time
* \*Anaysis to show which variables predict cancels by cancel reason

# Deliverables:

1. R Code
2. Report detailing results
3. Slide presentation of results