IST-687 Final Project

Project Owners (Individual):

Sharat Sripada

Analysis on data-set

- **Understanding variables:**
  - Satisfaction
  - Arrival & Departure Delay in minutes
  - Airline Names (to determine Airline popularity)
  - No. of other loyalty cards
  - Correlating Satisfaction, Popularity, Delays
- **Analysis from data-crunching:**
  - Delay (variables Arrival/Departure time)

  West Airways Inc. arrives on average 6.691395 minutes late

  FlyFast Airways Inc. arrives on average 42.83042 minutes late

  Southeast Airlines Co. arrives on average 19.20768 minutes late
  - Popular airliners (variable Airline Names)

  Cheapseats Airlines Inc. is the most used airline (20.06175 % of all passenger bookings)

  Cool&Young Airlines Inc. is the least used airline (0.9916159 % of all passenger bookings)

  Southeast Airlines Co. is at # 7 (7.373219 % of all passenger bookings)
  - Mean satisfaction (variable Satisfaction grouped per Airline)

  West Airways Inc. has highest average CSAT(3.486967)

  GoingNorth Airlines Inc. has lowest average CSAT(3.297194)

  Southeast Airlines Co. is at # 6 (average CSAT 3.396888)

- **Analysis from data-crunching (cont.):**
  - Applied Loyalty cards/discount coupons (variable No. of other Loyalty cards)

  Cheapseats Airlines Inc. has highest loyalty-cards offered/used (23209)

  Cool&Young Airlines Inc. has lowest loyalty-cards offered/used (1128)

  Southeast Airlines Co. is at # 7 (loyalty-cards offered/used 8439)

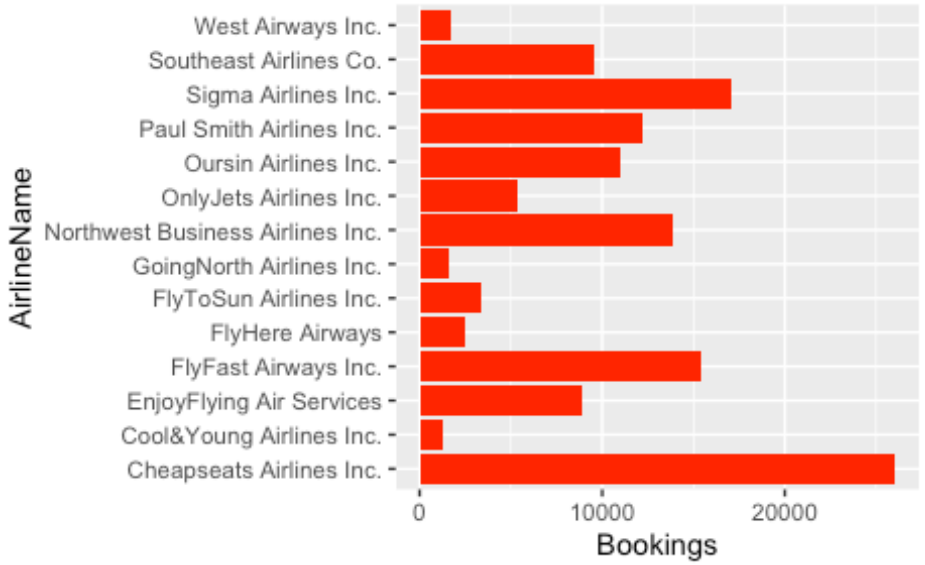
Business Questions

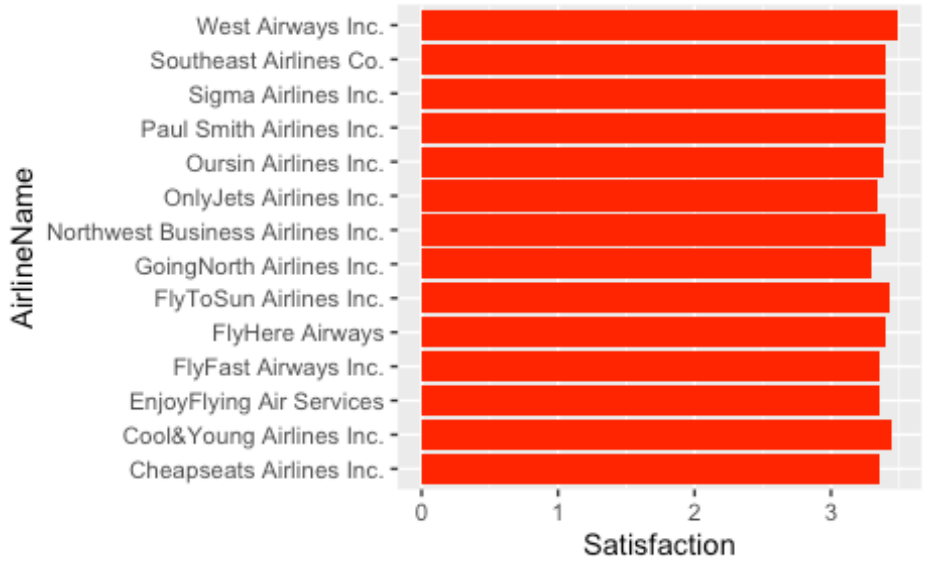
The intent when analyzing the airline booking data-set was to represent an airline and play the role of a Business Analyst or Consultant with specific function of Data-Scientist. In this context, it was determined from initial analysis that *Southeast Airlines Co.* was not the most preferred airline among passengers and that therefore would offer sufficient challenge and scope to analyze and make or show improvements.

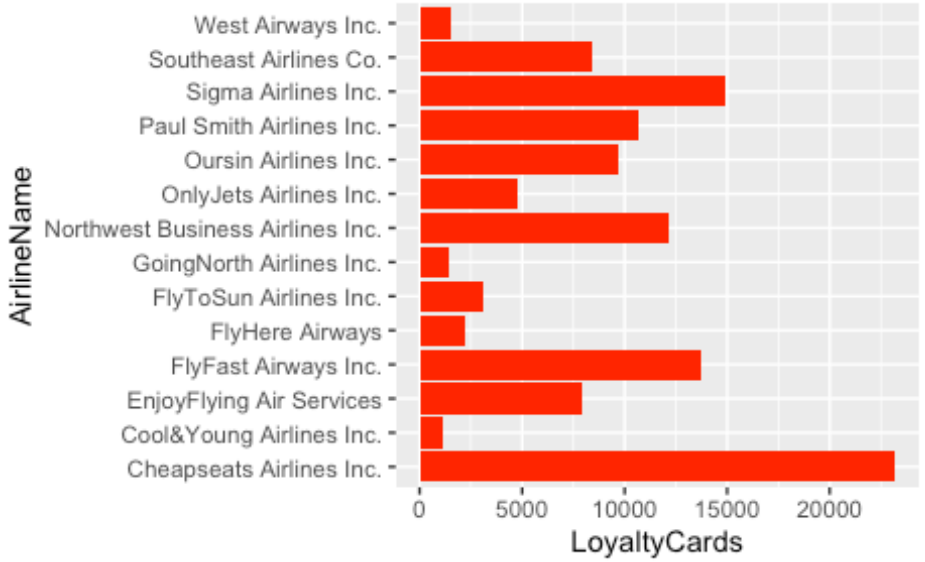
The analysis will broadly be based on the following:

1. Determine commonly top/worst performing airlines with respect to Airline booking
2. Key factors or variables likely to impact passenger satisfaction, airline bookings etc.
3. Using advanced prediction techniques provide consulting suggestions/improvements

Descriptive Stats – popularity, satisfaction, loyalty program







Data Modeling Techniques and Predictive Analysis

Linear Regression

- Low R-square values indicates low correlation & we are therefore unable to reject the NULL hypotheses – 'No correlation exists between Airline Names and No. of other Loyalty Cards.
- Low R-square values/correlation co-efficient between several combination of variables
- predictive analysis using Linear regression, factoring in results from Parsimonius model:
- All variables shortlisted – *11.44%*
- Variables as a result of Parsimonius Model/Step function did not yield very useful prediction as well at *11.48%*

Support vector machines

- **KSVM:**

```
library(kernlab)
cutpoint2_3 <- floor((2 * length(randindex) /3))
trainData <- df_models[randindex[1:cutpoint2_3],]
testData <- df_models[randindex[(cutpoint2_3 + 1):length(randindex)],]

ksvmoutput <- ksvm(`Airline Name`~., data=trainData,
                   kernel="rbfdot", #kernel function that projects the low dimensional problem into higher dimensional space
                   kpar="automatic", #params used to control radial function kernel(rbfdot)
                   C=10, #C -> cost of constraints
                   cross=10, #use 10 fold cross-validation in this model
                   prob.model=TRUE)
```

Prediction accuracy - 11.74%

- **SVM:**

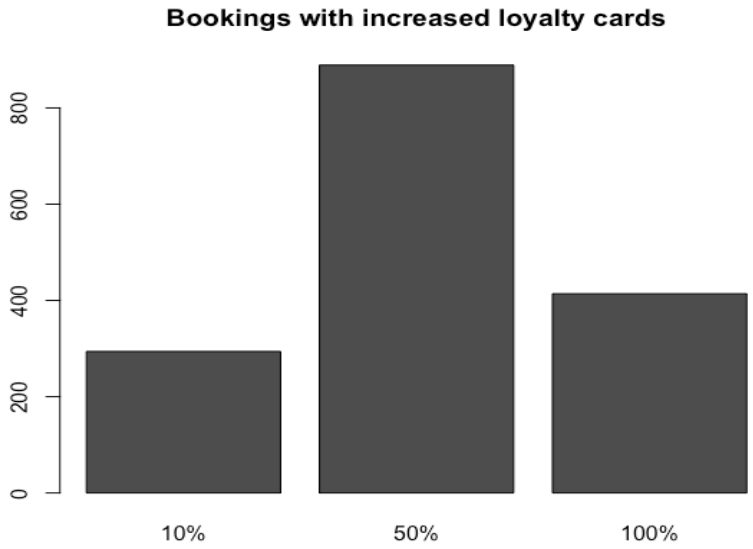
```
library(e1071)
svmoutput <- svm(`Airline Name`~., data=trainData,
            kernel="linear", #kernel function that projects the low dimensional problem into
higher dimensional space
            cross=10, #use 10 fold cross-validation in this model
            scale=FALSE)

svmpredict <- round(predict(svmoutput, test, type="response"))
svm_compTable <- data.frame(testData[,1], svmpredict)
colnames(svm_compTable) <- c('Test', 'Pred')

percentage_svm <- length(which(svm_compTable$Test ==
svm_compTable$Pred))/dim(svm_compTable)[1]
```

Prediction accuracy - 11.44%

Insights



Percentage increase in loyalty cards offered	Bookings per data set (in numbers)	Bookings per data set (in %)
10%	294	3.9%
50%	889	11.88%
100%	414	5.53%

As an experiment, 'No. of loyalty cards' variable was adjusted in the following manner for SouthEast Airlines Co.:

- If no cards were offered, offer one card at the least
- If cards were already offered then increase the number of loyalty cards by 10%, 50% and 100%

Conclusion

Clearly in my opinion variable 'No. of Loyalty cards' seems a mere distractor and cannot be used in any meaningful analysis. Also, that statement possibly holds some ground to the dataset in entirety. As a Data-Scientist or Business Analyst for Southeast Airlines Co. I would recommend getting more insightful data:

- Adding passenger first and last names can possibly help in determining how a passenger truly rated his/her experience when flying an airline provider and how that impacted flying again with the same airline in the future
- Specific satisfaction indices – related to delay and service