

|  |
| --- |
|  |
| A Data Science Project by  Doug Gerard, Tamora Long, Tyler Marsh, & Nicole Novakidis​  ​ |
| |  |  |  | | --- | --- | --- | | IST 687 Section 2019-0109 | Professor Gary Krudys | Group 4 – Team Ank | |



TABLE OF CONTENTS

[I. Introduction 2](#_Toc3740939)

[II. Business Questions 3](#_Toc3740940)

[III. Data Acquisition & Cleansing 5](#_Toc3740941)

[IV. Key Variables 6](#_Toc3740942)

[V. Descriptive Statistics & Visualizations 7](#_Toc3740943)

[**V.1 Southeast Flight Demographics by Gender** 7](#_Toc3740944)

[**V.2 Southeast Flight Demographics by Age** 9](#_Toc3740945)

[**V.3 Southeast Flight Demographics by Travel Type** 10](#_Toc3740946)

[**V.4 Southeast Flight Demographics by Status** 13](#_Toc3740947)

[**V.5 Satisfaction: Flight Distance vs. Flight Time** 15](#_Toc3740948)

[VI. Modeling Techniques and Predictions 17](#_Toc3740949)

[VII. Summary & Actionable Insights 19](#_Toc3740950)

[VIII. Appendix (R Code) 20](#_Toc3740951)

I. Introduction

Our firm, Ankylo Consulting, was assigned to be the acting consultants for our client, Southeast Airlines. We were provided a dataset, a survey of customers flying within the United States from January 1, 2014 through March 31, 2014. Our goal was to predict those customers with low satisfaction, analyze why this might be, and determine what actions to take to gratify them in order to gain customer loyalty.

The data set contained approximately 130,000 participants, with 28 fields. This included attributes such as age, gender, price sensitivity, status, delays, etc., for both Southeast Airline, and some of its direct competitors.

The following report details the analysis found by Ankylo Consulting based on the dataset we received. Important criteria for gauging customer satisfaction in the survey was the rating. With the provided ratings, we narrowed down which customers we would analyze based on the following criteria:

* **Rating of 6-10:** These customers are mostly, or completely, satisfied and would continue to fly on Southeast Airlines.
* **Rating of 5:** These customers are neutral and may not have much of a preference of flying with Southeast Airlines or a competitor airline.
* **Rating of 1-4:** These customers are unsatisfied and are looking for improvements to be made in order to become loyal customers of Southeast Airlines.

The report will provide more information to support our analysis including statistics, considered variables, and visualizations. In the closing of this document, we will provide a summary of thoughts and advise on actionable steps that Southeast Airlines should perform to improve overall customer satisfaction.

II. Business Questions

The following business questions were developed to help analyze the current satisfaction levels of Southeast Airlines customers.

1. **How is the customer satisfaction overall?**

* Overall satisfaction of all airlines has a rating of *4.98 of 10*
* Southeast Airlines has a satisfaction rating is *5.00 of 10*

1. **What are the measurement criteria of a disappointed customer?**
   * We have determined that "unsatisfied" customers submit a rating of *1,2,3, or 4*
2. **What is the profile of a disappointed customer?**

When analyzing this business question, some of the factors considered included: age, gender, flight tenure, travel purpose, flight class, etc. Here are findings:

* + Many disappointed customers have a general profile with the age of 50.
  + A large amount of disappointed customers’ gender is female
  + Many disappointed customers fly for personal travel versus the business or mileage travel types
  + Many disappointed customers are low-level, “blue status”, in the mileage program.

1. **What is causing the disappointment?**

The main factors that seem to be causing disappointment include customers flying for personal traveling, flights that have an arrival delay, customers that have a status of “Blue” and flight cancellations. Here are some possible explanations:

* Flying for personal reasons means that the customer is paying out of their own pockets, as opposed to a business paid trip. They are expecting more for their dollar.
* Flights that have an arrival delay, or an actual cancellation, is a cause for disappointment. Leading customers to miss deadlines or planned events.
* Having a “blue status” means less perks, like seat upgrades and free checked bags. These customers travel with less comfort than those with a higher status level.

1. **Are there correlations between multiple factors? If so, which correlations are most impactful?**

**Correlations between our independent varibles:**

**We tested the predictors Variance inflation Factor, VIF, and two sets of variables were correlated with one another. It appears that the company and city a passenger are flying to actually share much of the same variance in our models. Flight time and distance also were correlated in our model. This is not surprised considering the farther a flight goes the longer it will be contributing to the same issues experienced by passengers.**

**Airline.Name 65.823872 13 1.174729**

**dataSet$Destination.City 71.216581 295 1.007256**

**dataSet$Flight.time.in.minutes 18.645788 1 4.318077**

**dataSet$Flight.Distance 18.487034 1 4.299655**

1. **What are the other factors aside from the actual flight that is causing this disappointment?**

When analyzing this business question, some of the factors considered included: options and quality of the food court, airport shopping, loyalty program, airline comparison, tenure in flying, travel date, etc. Here are findings:

* + The amount of money customers spend on shopping and food plays a small role in the satisfaction of customers. According to our linear models, over 90% of Southeast’s customers shopping and amount spent on eating accounted for less than 1% of their satisfaction.
  + The years a customer spent flying accounts for a small percentage of their satisfaction overall as well. The mean number of years a person has been flying subtracts .2 from their satisfaction on a scale of 1-10
  + Price sensitivity seems to play a role in customer dissatisfaction. For each additional point of price sensitivity satisfaction seems to go down by 3.81%. The difference between the least price sensitive customer and most seems to be a little over 10%. This could be the difference of a neutral customer to negative and positive to neutral.

**7. What are ideas to gain their satisfaction?**

A few of our ideas to gain customer satisfaction are (see section VII for more detail):

* Lower the price for flights with short distances
* Provide a few more perks for customers with blue status
* Provide more comfort

III. Data Acquisition & Cleansing

After loading the data set in R Studio, our first course of action was to clean up the dataset. The following decisions were made:

* There were 3 columns with NAs: Departure Delay, Arrival Delay & Flight time. We decided to replace the NAs with the mean of the column.
* Binary columns were created to represent each of the discrete data items.
* Removed all leading and trailing white spaces.
* Changed column names to lowercase in case there was any "dirty data."
* Agreed that "unsatisfied" customers would be those with a rating of 1-4, with 5 being neutral, and 6-10 being "happy."

IV. Key Variables

After some initial analysis, we determined which of the 28 variables we should focus on. The following table gives a brief outline on the decided upon key variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable**​ | **Type**​ | **Initial Indications**​ | | |
| Gender​ | Discrete​ | Initially indicates females are a larger share of unhappy customers​ | | |
| Age​ | Continuous​ | Initially indicates those over 60 are nearly 40% of complaints​ | | |
| Class​ | Discrete​ | Economy passengers are 82% of complaints​ | | |
| Mileage Status​ | Discrete​ | Blue level accounts for 82% of complaints​ | | |
| Eating /Drinking​ at the Airport | Continuous​ | 67% complaints are from individuals spending $20 or more on food/drink ​ | | |
| Price Sensitivity​ | Discrete​ | Most price sensitive customers account for 64% of complaints​ | | |
| Travel Type​ | Discrete​ | Personal travel accounts for 57% of complaints​ | | |
| Loyalty Card Count​ | Discrete​ | 58% of unhappy customers have no loyalty card​ | | |
| Departure Delay​ | Continuous​ | 59% of complaints are associated with flights with no departure delay​ | | |
| Arrival Delay​ | Discrete​ | 59% of complaints are for flights without arrival delay​ | | |
| Percent travel on Other Airlines​ | Discrete​ | 68% of complaints by those that travel 10-1% on other airlines​ | | |
| Airline​ | Discrete​ | 25% are from Cheapseats airline​ | | |
| Flight Time​ | Continuous​ | Flights between 20 and 120 minutes have most complaints​ | | |
| **Variable** | | | **Type** | **Initial indications** |

V. Descriptive Statistics & Visualizations

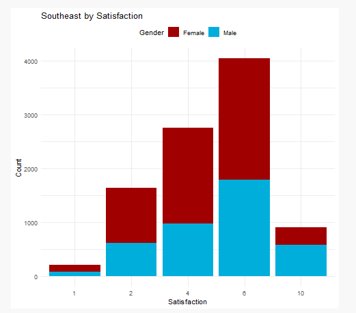
Again, we used the rating criteria mentioned before against our key variables to analyze the provided dataset. Here is a reminder of the ratings:

* **Rating of 6-10:** These customers are mostly, or completely, satisfied and would continue to fly on Southeast Airlines.
* **Rating of 5:** These customers are neutral and may not have much of a preference of flying with Southeast Airlines or a competitor airline.
* **Rating of 1-4:** These customers are unsatisfied and are looking for improvements to be made in order to become loyal customers of Southeast Airlines.

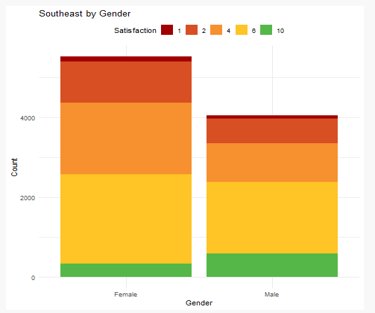
We will display our findings visually.

**V.1 Southeast Flight Demographics by Gender**

Southeast has more female travelers than male travelers

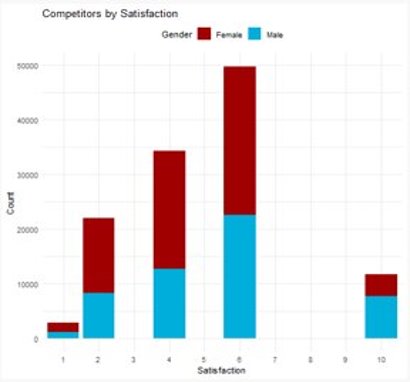


Females are more unhappy at ~50% rating of 6 or higher vs. males at ~56% rating of 6 or higher



**How does Southeast compare to competitors by gender?**

Southeast's competitors also serve more female travelers

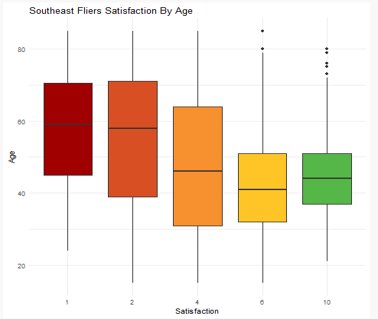


**V.2 Southeast Flight Demographics by Age**

The average age of Southeast travelers are between the ages of 30-70 years old ​

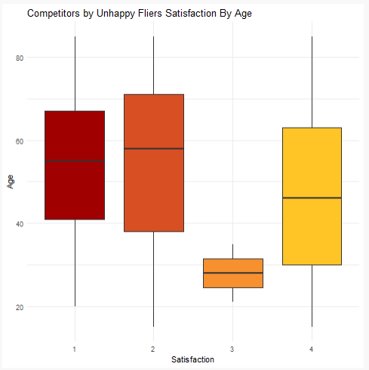
​

The most unhappy fliers, with a rating of 1 or 2, have a median age of around 57 years old

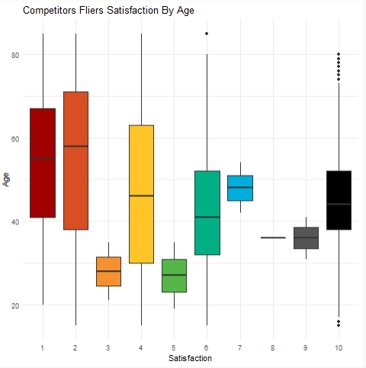


**How does Southeast compare to competitors in Age?**

The most unhappy fliers of Southeast's competitors, with a rating of 1 or 2, have a median age of around 57 years old

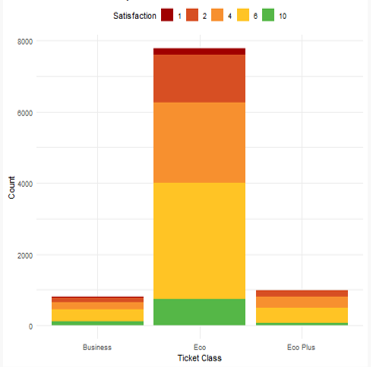


Southeast Airlines generally serve the same demographics as its competitors by age

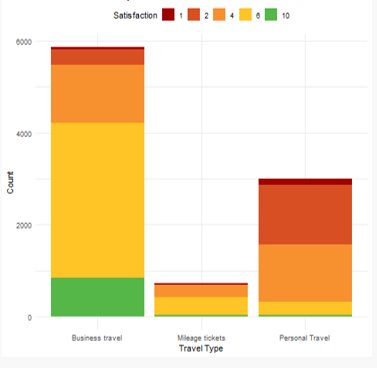
****

**V.3 Southeast Flight Demographics by Travel Type**

Most travelers fly in economy ticket class

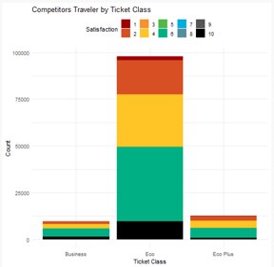


Most travelers fly for business purposes

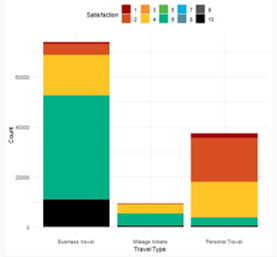


**How does Southeast compare to competitors in Travel Type?**

Most competitors’ travelers also fly in economy ticket class

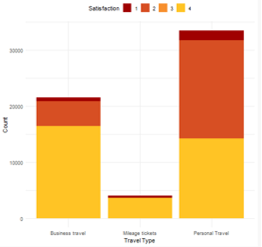


Most competitors’ travelers also fly for business purposes



**How does Southeast compare to competitors in Travel Type (cont.)?**

The most unhappy travelers of Southeast's competitors fly for personal travel purposes as well



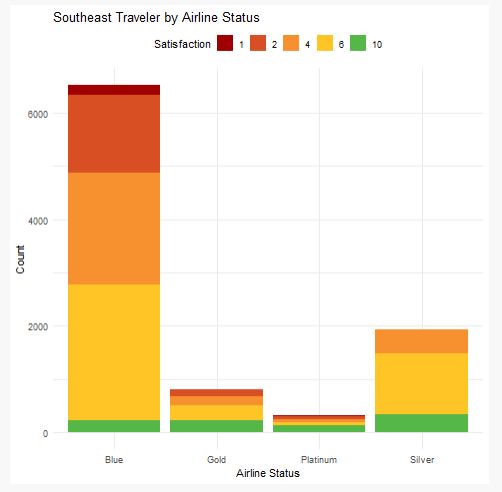
**V.4 Southeast Flight Demographics by Status**

Most of Southeast's travelers are Blue status in the rewards program

* 60% of Southeast's travelers that are unhappy, with a rating of 4 or less, are Blue status

Southeast's most happy travelers are at Silver status

* 79% of travelers with Silver status are happy with their flight experience, rating the experience at 6 or above

****

**How does Southeast compare to competitors by Status?**

Most of Southeast's competitors have travelers with Blue status in the rewards program

* 57% of travelers that are unhappy, with a rating of 4 or less, are Blue status

Their happiest travelers are also Silver status

* 76% of travelers with Silver status are happy with their flight experience, rating the experience at 6 or above

A screenshot of a cell phone

Description generated with high confidence

**V.5 Satisfaction: Flight Distance vs. Flight Time**

The most unhappy travelers are on shorter flights

Flight travel times are longer than competitors for same distances

A close up of a map

Description generated with high confidence

Longer flights show a trend for happier travelers

A close up of a map

Description generated with very high confidence

VI. Modeling Techniques and Predictions

We decided to use a few modeling techniques to see what causes customers to be satisfied and unsatisfied. We modeled the data with multiple linear regression models and support vector machines. Here are the techniques we used when considering which model to use in the prediction of satisfied/unsatisfied customers.

* **Multiple Linear Regression**: is a technique that explores the relationship between a dependent variable and multiple independent variables. It has many practical applications and is commonly used in prediction and quantifying the strength of the relationship between predictors and the outcome.
* **Akaike Information Criterion (AIC)**: is an estimate of the quality of a statistical model. It is commonly used to compare different models as means of selecting which model to use.
* **Support Vector Machine**: is a form of machine learning with associated learning algorithms that are used for regression and classification. In classification a Support Vector Machine (SVM) takes data marked as belonging to categories then builds an algorithm to predict these categories given the data with future data.

Predicting Unhappy Southeast customers using Multiple Linear Regression

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.1009295 0.0935127 33.161 < 2e-16 \*\*\*

Age -0.0045048 0.0007045 -6.394 1.77e-10 \*\*\*

code.BlueStatus 0.7215486 0.0864087 8.350 < 2e-16 \*\*\*

code.SilverStatus 1.7645876 0.0955172 18.474 < 2e-16 \*\*\*

code.GoldStatus 0.7994093 0.1000034 7.994 1.64e-15 \*\*\*

code.MileageTravel 0.3948433 0.0587759 6.718 2.07e-11 \*\*\*

code.PersonalTravel -0.5919265 0.0302817 -19.547 < 2e-16 \*\*\*

code.arrivalDelay -0.6097602 0.0285832 -21.333 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9143 on 4609 degrees of freedom

Multiple R-squared: 0.2527, Adjusted R-squared: 0.2515

F-statistic: 222.6 on 7 and 4609 DF, p-value: < 2.2e-16

Step: Predicting Unhappy Southeast customers using Akaike Information

Criterion for predictions

AIC=-836.97

code.Satisfaction ~ Age + code.female + code.BlueStatus + code.SilverStatus +

code.GoldStatus + code.EcoPLusClass + code.MileageTravel +

code.PersonalTravel + code.Sensitivity.VeryHigh + code.Sensitivity.High +

code.Sensitivity.Medium + code.arrivalDelay + No.of.Flights.p.a. +

Eating.and.Drinking.at.Airport

Df Sum of Sq RSS AIC

<none> 3826.6 -836.97

- code.Sensitivity.Medium 1 2.36 3828.9 -836.12

- No.of.Flights.p.a. 1 2.60 3829.2 -835.83

- code.Sensitivity.High 1 3.70 3830.3 -834.51

- code.Sensitivity.VeryHigh 1 3.86 3830.4 -834.32

- Eating.and.Drinking.at.Airport 1 4.27 3830.8 -833.82

- code.EcoPLusClass 1 4.97 3831.5 -832.98

- code.female 1 8.48 3835.0 -828.75

- Age 1 28.54 3855.1 -804.66

- code.MileageTravel 1 35.57 3862.1 -796.25

- code.BlueStatus 1 49.11 3875.7 -780.10

- code.GoldStatus 1 49.47 3876.0 -779.66

- code.SilverStatus 1 274.87 4101.4 -518.69

- code.PersonalTravel 1 319.69 4146.3 -468.52

- code.arrivalDelay 1 381.14 4207.7 -400.59

VII. Summary & Actionable Insights

(Summary Paragraph)

Through our research, Ankylo Consulting found that many of Southeast Airlines’ unhappy customers are price-sensitive. This, coupled with our findings that unhappy customers are also those on shorter flights, may mean that the price is too high for flying short distances. Southeast airlines may want to lower the price for short distances, with a possible algorithm based on the number of miles flown.

Many unhappy customers are in the low-level blue status. It is likely that these customers are not flying often enough to gain a higher status, but an additional perk can be made in order to gain loyalty so that Southeast Airlines is flown anytime they do travel. Ideas for perks are a percentage discount for multiple bags checked for the short-term, and extra miles given for certain flights, if not all, so that they can enter the next status level faster. The price for longer flights can be left alone, or even increased slightly, to offset.

The average age of an unhappy traveler, 57, is also something to consider. It is possible that there is a comfort factor involved. Seats are much smaller than they used to be and sitting for even short periods of time may be more bothersome to a flyer in that age range. Ideas for comfort are making extra cushions, such as head rests and lumbar support, available.

VIII. Appendix (R Code)

# R Source Summary

In approach the data set the code take the general approach of implementing functions that can repeatedly apply analysis and plot generation on any set or subset of the data. This approach leaves broad latitude to analyze the data as plots, regression and analysis can be done on the data with a few function calls. The approach was adopted early on as the team need to view the data in a number of ways for example Southeast’s customers, the industry as a whole, Southeast’s competitors, satisfied customers, unsatisfied customers and any other subset of the data.

While much of the generated work yielded no useful information taken in all it allowed the team to quickly ‘get a feel’ for the data to home in on patterns.

## Packages Used

MASS, Support Functions and Datasets for Venables and Ripley's MASS, v7.3.51.1

utils, The R Utility Package, v3.5.2

ggplot2, Create Elegant Data Visualizations Using the Grammar of Graphics, v3.1.0

ggcorplot, Visualization of a Correlation Matrix using ggplot2, v0.1.2

RColorBrewer, Color Brewer Palettes, v1.1.2

## Function Summaries

### AnkylosaurusPal

Provided a palette to use for plots by name: main, all, cool, hot, mixed, grey, binary

### ScaleColorAnkylosaurus

Returns a scaled/gradient color palette based on defined color palettes from *AnkylosaurusPal* to use for plots where a gradient was needed.

### Trim

Remove white same from with end of a string

### ToLowerCase

Return a string in all lower case

### TrimLowerCase

Time a string and return in in lowercase

### Map

Used to map discrete data into a integer value that allowed for easy manipulation. For example, mapping the customer class from Blue, Silver Gold and Platinum to 1, 2, 3, and 4 respectively. Deliberately used to ensure all data matched an expected pattern to allow for easy comparison.

### ChiSquareTest

Implements a ChiSquare test on discrete data, printing the results.

### NADescription

Prints information about the passed in data set and the number of rows with invalid ‘NA’ data values, and details which columns contained the missing values.

### DescribeData

Print a summary of the passed in data frame and describes the ‘NA’ value distribution amount the data frame’s columns.

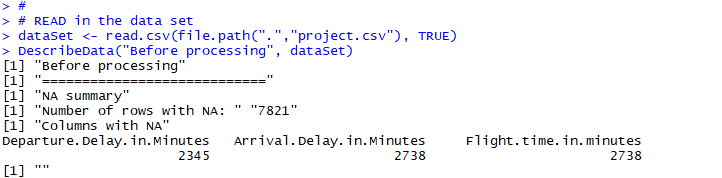
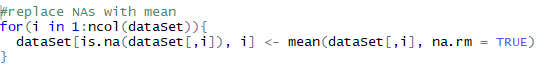
### PlotSet

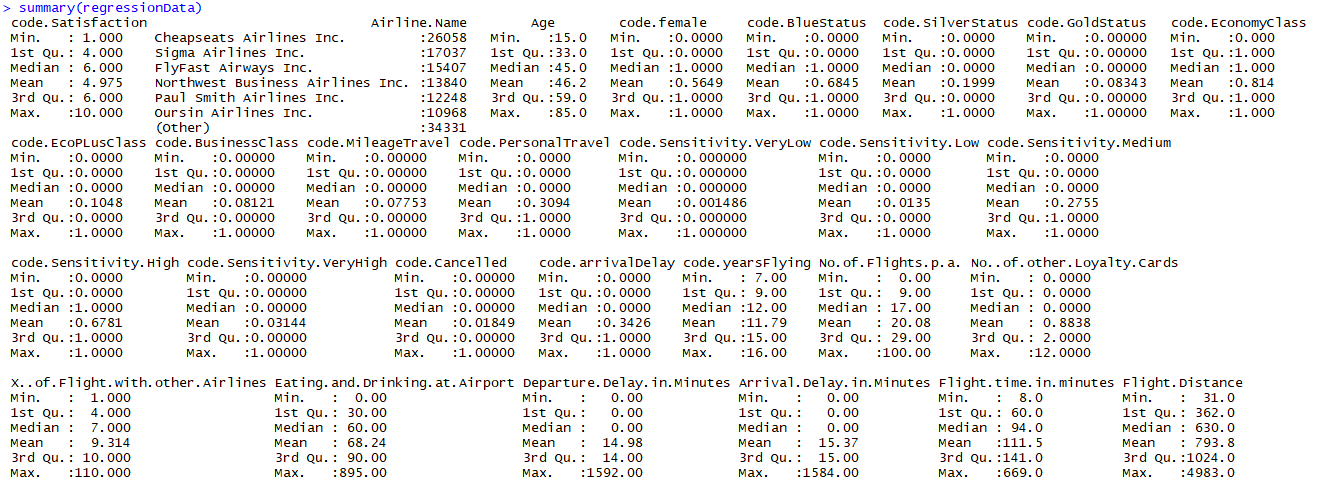
Generate a variety of plots for a data frame that is passed into the function.

### RelationTests

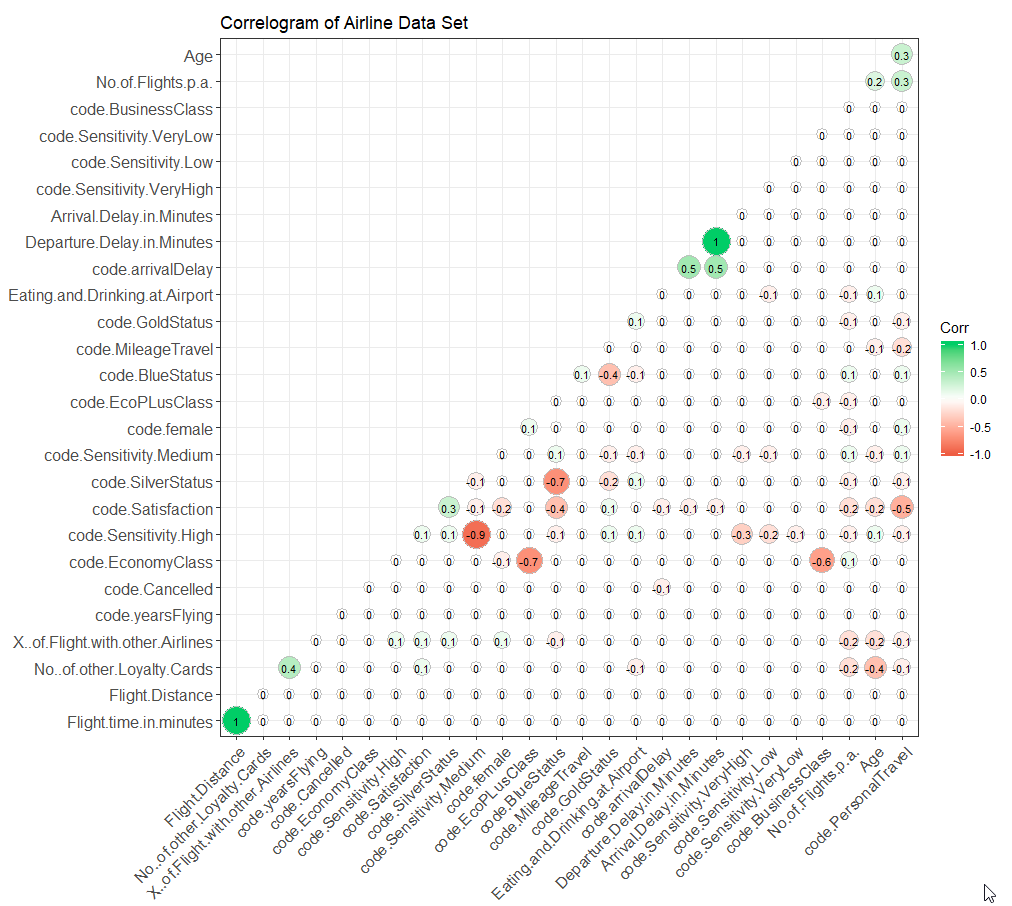
Executes tests and generates plots to test hypotheses.

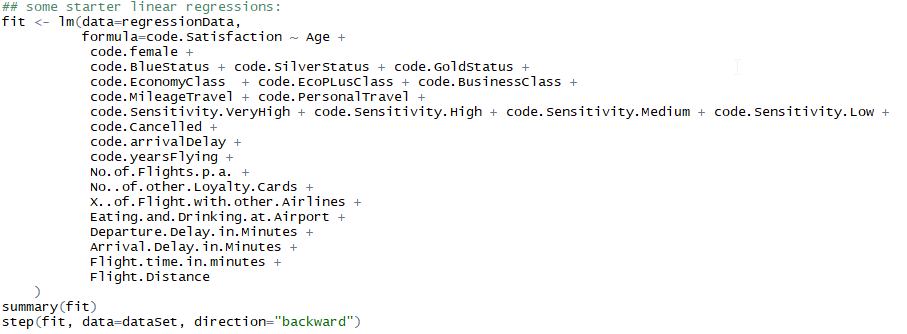
## Data manipulation

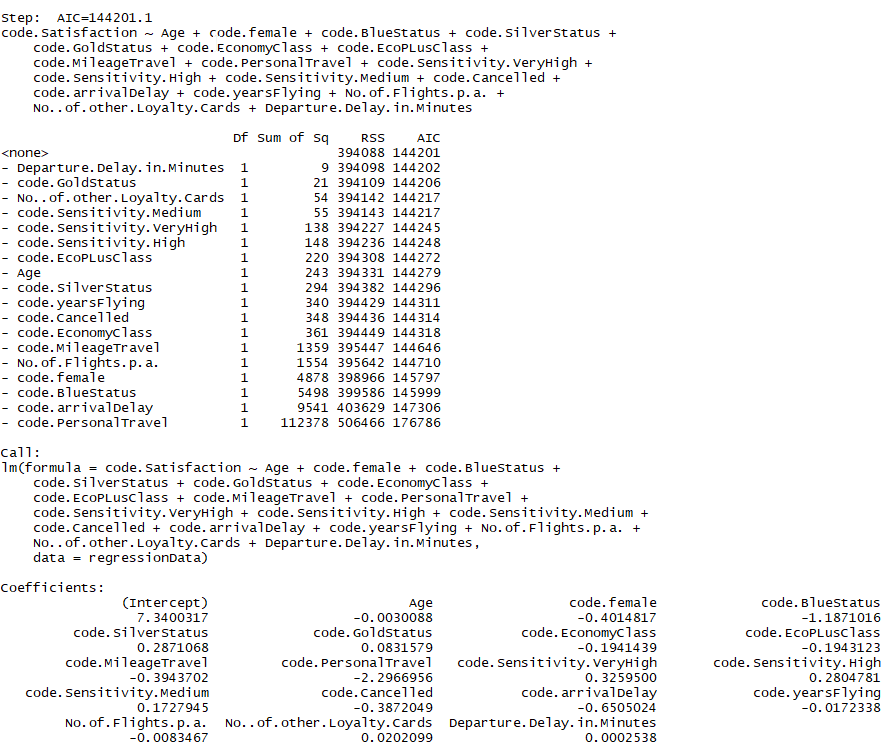
* Data was manipulated to ensure where data should have discrete values (i.e. Customer Satisfaction) it was rounded down to the nearest value for entries that has non-standard values such as ‘4.5’
* Data was loaded and prior to processing it was discover that 7821 rows contained one or more data items with not value as indicated by ‘NA’
  + 
* NA were removed from the data and as the three columns that contained the NAs were for continuous data the mean of the entire data set was substituted for the NA in the appropriate column.
  + 
* Data was also divided into several different subsets:
  + All unhappy customers (satisfaction < 5)
  + All happy customers (satisfaction > 5)
  + Southeast’s customers
  + Southeast’s unhappy customers
  + Southeast’s happy customers
  + The competition’s customers
  + The competition’s unhappy customers
  + The competition’s happy customers
* A correlogram was generated for the entire set of data used in the regression test: 
  + Summary:



* + Correlogram:





* The final regression with the best AIC score:
  + 

## R Source



