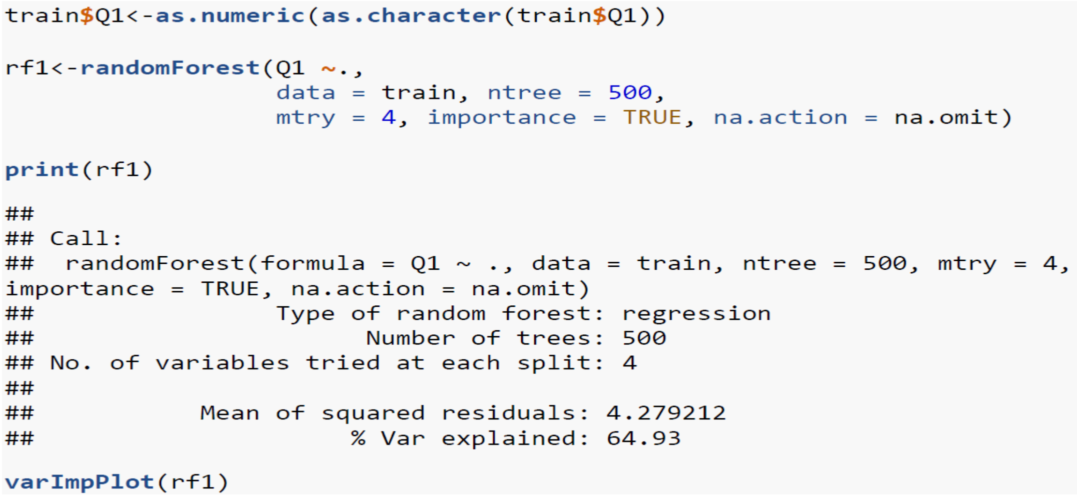
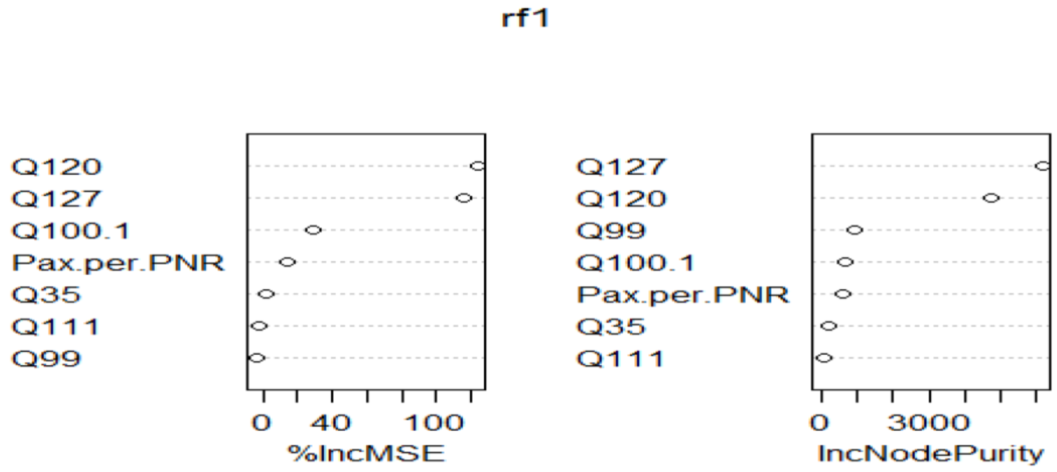
Guest Satisfaction Survey Key Factors

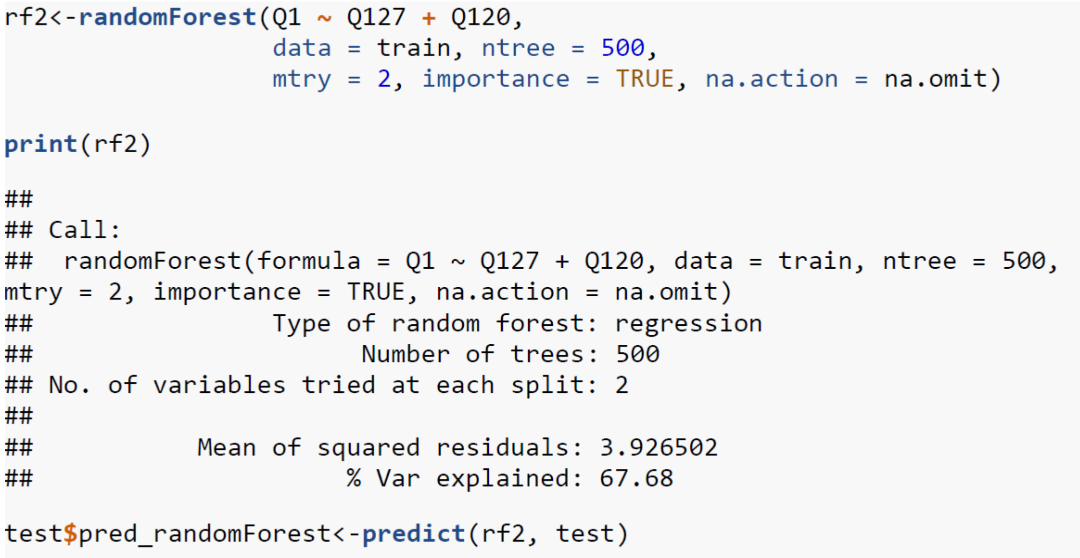
- Process:
    - Understand the airlines industry.
    - Understand/frame the business problem.
    - Explore the data & potential key factors with key business question in mind.
    - Look at all levels of the variables.
    - Transform all 'ordinal' columns into 'numerical columns'
    - Run histograms on transformed columns to see/understand data patterns.
    - Treat missing values by deleting columns that contained > 77% null values
    - Train/validate a Random Forest Regression Model
    - Look at the Variable Importance Plot in R for variable selection to predict the response variable Q1.
    - Upon that, select the top/three factors that best explain the response variable

Random Forest Model Technical Analysis

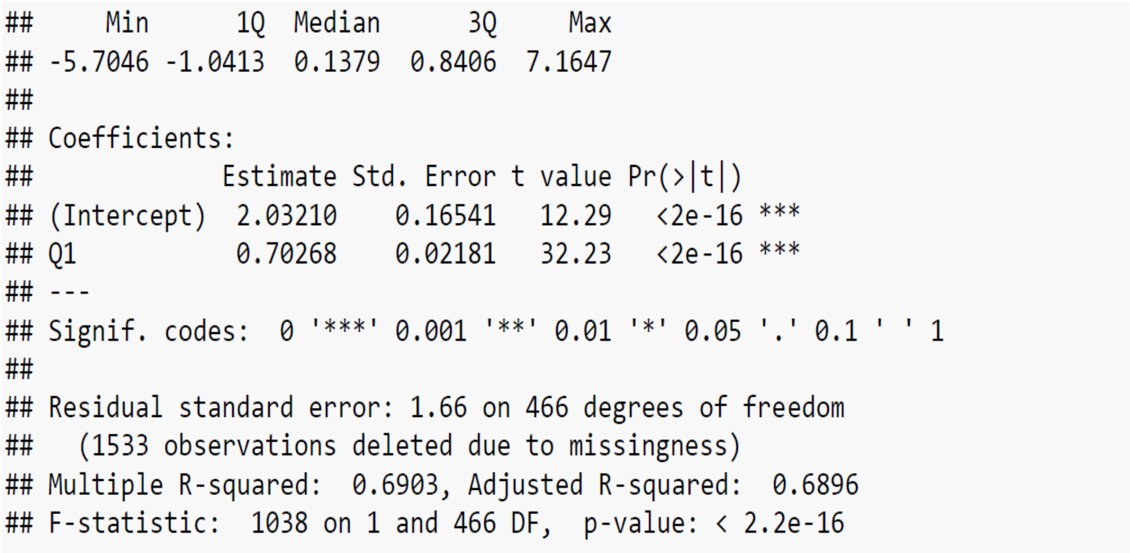


```
train$Q1<-as.numeric(as.character(train$Q1))

rf1<-randomForest(Q1 ~.,
                  data = train, ntree = 500,
                  mtry = 4, importance = TRUE, na.action = na.omit)

print(rf1)

##
## Call:
##  randomForest(formula = Q1 ~ ., data = train, ntree = 500, mtry = 4,
importance = TRUE, na.action = na.omit)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 4.279212
##                     % Var explained: 64.93

varImpPlot(rf1)
```





```
rf2<-randomForest(Q1 ~ Q127 + Q120,
                  data = train, ntree = 500,
                  mtry = 2, importance = TRUE, na.action = na.omit)

print(rf2)

##
## Call:
##  randomForest(formula = Q1 ~ Q127 + Q120, data = train, ntree = 500,
mtry = 2, importance = TRUE, na.action = na.omit)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 3.926502
##                     % Var explained: 67.68

test$pred_randomForest<-predict(rf2, test)
```



```
##     Min      1Q  Median      3Q     Max
## -5.7046 -1.0413  0.1379  0.8406  7.1647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.03210    0.16541   12.29   <2e-16 ***
## Q1           0.70268    0.02181   32.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.66 on 466 degrees of freedom
##   (1533 observations deleted due to missingness)
## Multiple R-squared:  0.6903, Adjusted R-squared:  0.6896
## F-statistic:  1038 on 1 and 466 DF,  p-value: < 2.2e-16
```

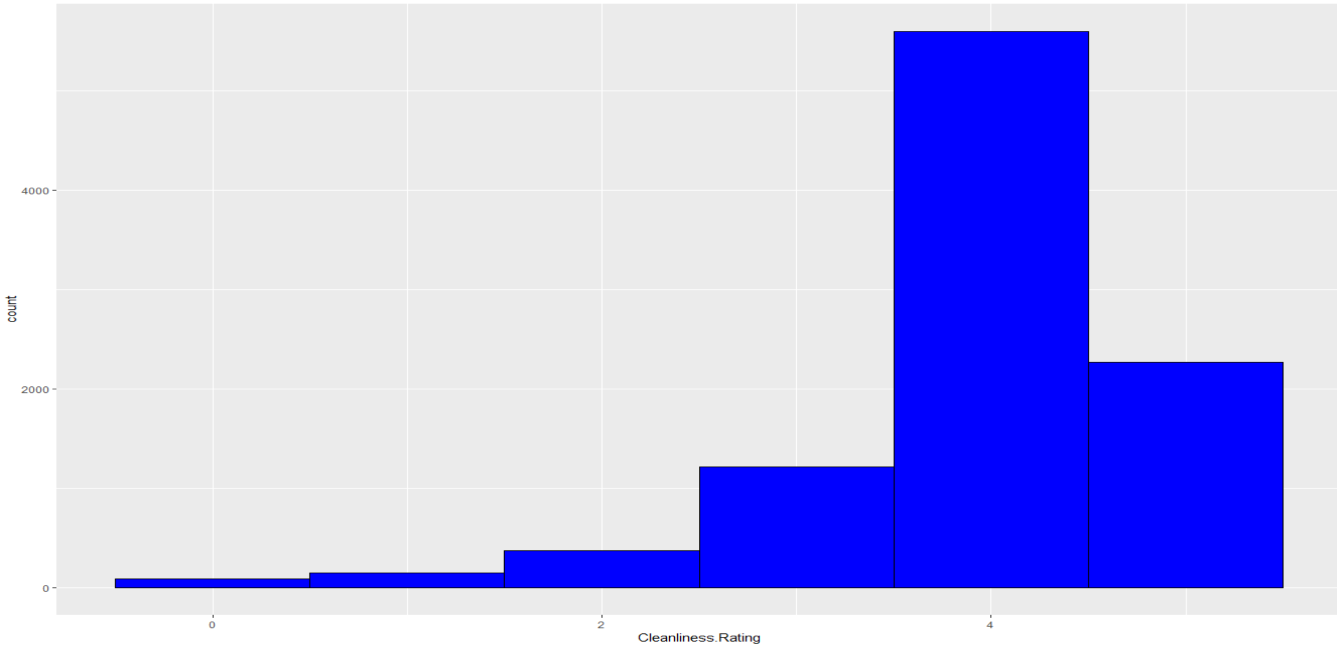
Random Forest Results – Key Drivers Interpretation

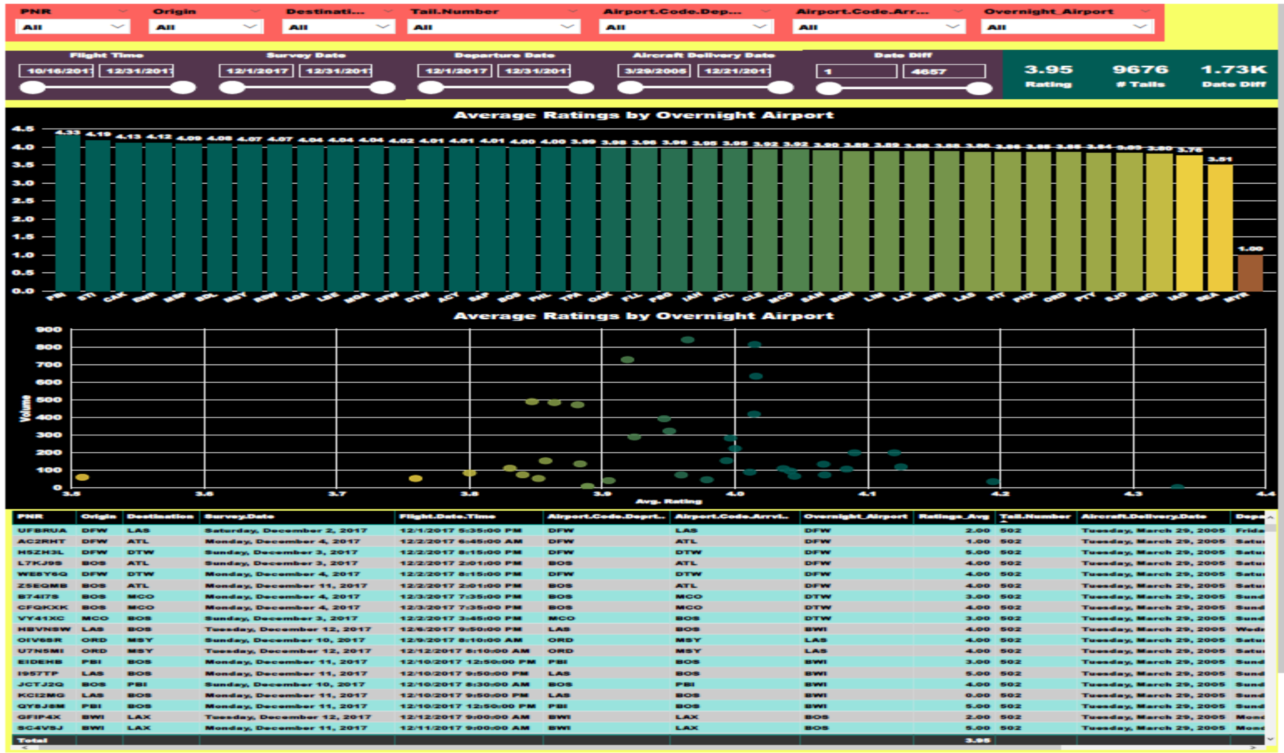
- Potential Key Drivers: Q127 and Q120
- Explain roughly 67.68% of the response variable ‘Q1’ among the training data set.
- Then predict/validate on the test data set, Adjusted R-square = 0.6896.
- Fairly consistent measures on both training/test set
- Overall Goal: Find the least number of key factors that explain the highest variance of the response variable to avoid multicollinearity

Aircraft Cabin Cleanliness Ratings Analysis

Tools Used: Excel, R (ggplot2), PowerBI

1. Copy paste Excel Test file into individual CSV Files and import them into R. Do basic data exploration.
2. Left outer join Tail Number Data so that the specified Tail Numbers would have a cabin cleanliness rating.
3. Concatenate Tail Number and Departure Date in Overnight Station Data for a unique ID.
4. Concatenate Tail Number and Departure Date in Test so you can join Overnight Station Data on Test Data.
5. Get rid of Tail Number, Departure Date in Overnight Station Data to get rid of duplicate columns (when they join)
6. Change the name of "Airport Code Deport" to "Overnight Airport "in Overnight Station Data
7. Left outer join Overnight Station Data on Test to get Overnight Airport matched up with cleanliness rating in original Test Data.
8. Left outer join Airport Delivery Dates on Test to get Aircraft Delivery Dates matched up with cleanliness rating in original Test Data.
9. Convert cleanliness rating where 0 = filthy and 5 = spotless.
10. Plot histogram to for data exploration using ggplot2.
11. Write to csv and do further analysis in PowerBI.





PNR: All
Origin: All
Destinati...: All
Tail.Number: All
Airport.Code.Dep...: All
Airport.Code.Arr...: All
Overnight_Airport: All

Flight Time: 10/16/2017 – 12/31/2017
Survey Date: 12/1/2017 – 12/31/2017
Departure Date: 12/1/2017 – 12/31/2017
Aircraft Delivery Date: 3/29/2005 – 12/21/2017
Date Diff: 1 – 4657

3.95 Rating
9676 # Tails
1.73K Date Diff

PNR	Origin	Destination	Survey.Date	Flight.Date.Time	Airport.Code.Deprt..	Airport.Code.Arrvl..	Overnight_Airport	Ratings_Avg	Tail.Number	Aircraft.Delivery.Date	Depa
UFBRUA	DFW	LAS	Saturday, December 2, 2017	12/1/2017 5:35:00 PM	DFW	LAS	DFW	2.00	502	Tuesday, March 29, 2005	Frida
AC2RHT	DFW	ATL	Monday, December 4, 2017	12/2/2017 6:45:00 AM	DFW	ATL	DFW	1.00	502	Tuesday, March 29, 2005	Satur
H5ZH3L	DFW	DTW	Sunday, December 3, 2017	12/2/2017 8:15:00 PM	DFW	DTW	DFW	5.00	502	Tuesday, March 29, 2005	Satur
L7KJ9S	BOS	ATL	Sunday, December 3, 2017	12/2/2017 2:01:00 PM	BOS	ATL	DFW	4.00	502	Tuesday, March 29, 2005	Satur
WE8Y6Q	DFW	DTW	Monday, December 4, 2017	12/2/2017 8:15:00 PM	DFW	DTW	DFW	4.00	502	Tuesday, March 29, 2005	Satur
Z5EQMB	BOS	ATL	Monday, December 11, 2017	12/2/2017 2:01:00 PM	BOS	ATL	DFW	4.00	502	Tuesday, March 29, 2005	Satur
B74I7S	BOS	MCO	Monday, December 4, 2017	12/3/2017 7:35:00 PM	BOS	MCO	DTW	3.00	502	Tuesday, March 29, 2005	Sund
CFQKXK	BOS	MCO	Monday, December 4, 2017	12/3/2017 7:35:00 PM	BOS	MCO	DTW	4.00	502	Tuesday, March 29, 2005	Sund
VY41XC	MCO	BOS	Sunday, December 3, 2017	12/2/2017 3:45:00 PM	MCO	BOS	DTW	3.00	502	Tuesday, March 29, 2005	Sund
HBVNSW	LAS	BOS	Tuesday, December 12, 2017	12/6/2017 9:50:00 PM	LAS	BOS	BWI	4.00	502	Tuesday, March 29, 2005	Wedr
OIV6SR	ORD	MSY	Sunday, December 10, 2017	12/9/2017 8:10:00 AM	ORD	MSY	LAS	4.00	502	Tuesday, March 29, 2005	Satur
U7N5MI	ORD	MSY	Tuesday, December 12, 2017	12/12/2017 8:10:00 AM	ORD	MSY	LAS	4.00	502	Tuesday, March 29, 2005	Satur
EIDEHB	PBI	BOS	Monday, December 11, 2017	12/10/2017 12:50:00 PM	PBI	BOS	BWI	3.00	502	Tuesday, March 29, 2005	Sund
I957TP	LAS	BOS	Monday, December 11, 2017	12/10/2017 9:50:00 PM	LAS	BOS	BWI	5.00	502	Tuesday, March 29, 2005	Sund
JCTJ2Q	BOS	PBI	Sunday, December 10, 2017	12/10/2017 8:30:00 AM	BOS	PBI	BWI	4.00	502	Tuesday, March 29, 2005	Sund
KCI2MG	LAS	BOS	Monday, December 11, 2017	12/10/2017 9:50:00 PM	LAS	BOS	BWI	0.00	502	Tuesday, March 29, 2005	Sund
QY8J8M	PBI	BOS	Monday, December 11, 2017	12/10/2017 12:50:00 PM	PBI	BOS	BWI	5.00	502	Tuesday, March 29, 2005	Sund
GFIP4X	BWI	LAX	Tuesday, December 12, 2017	12/12/2017 9:00:00 AM	BWI	LAX	BOS	2.00	502	Tuesday, March 29, 2005	Mond
SC4VSJ	BWI	LAX	Monday, December 11, 2017	12/11/2017 9:00:00 AM	BWI	LAX	BOS	5.00	502	Tuesday, March 29, 2005	Mond
Total								3.95			

Key Questions

Which airports are the best at cleaning aircraft?

PBI	STI	CAK	EWR	MSP	BDL	MSY	RSW	LGA	LBE
4.33	4.19	4.13	4.12	4.09	4.08	4.07	4.07	4.04	4.04

Do certain airports impact the aircraft cleanliness ratings more than others?

Yes, the four high-volume airports with the lowest aircraft cleaning ratings are MCO (3.92), ORD (3.85), LAS (3.86), BWI (3.88)

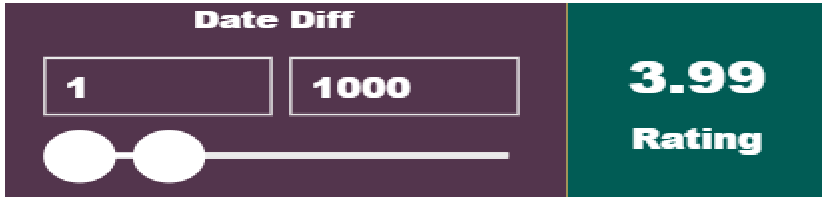
What other factors could drive the cleanliness ratings to be low?

Other factors would be time difference between the Aircraft Delivery Date and Departure Date.

Date Difference (in Days) between 0 Days and 1000 Days: 3.99 Rating

Date Difference (in Days) between 1000 Days To Max: 3.91 Rating

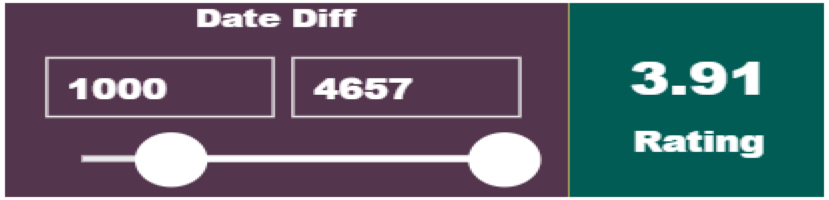
Obviously, further data analysis would allow us to look at certain thresholds with the worst ratings.



Date Diff

1 | 1000

3.99 Rating



Date Diff

1000 | 4657

3.91 Rating