MATH 324 Final Project

Amy Cao, Amy Folkestad, Anker Hojgaard and Jaxson Stathis

1. Project Description

Should be in paragraph form (not bullets). Written for a 3rd party client that may not have a strong statistical background. As part of your description, consider the following:

Gold Mine Resorts is looking to understand the predictors for guests who cancel their reservations.The industry average for cancellation rates is twenty percent; however Gold Mine Resorts experienced a 32.8% cancellation rate for reservations between 2015 and 2018. Cancelled reservations represent a total of $4.3 million dollars in lost revenue for this time frame.

- *Description of the dataset*

The dataset includes 36,285 rows of booking data, spanning the course of four years (2015-2018). Data collected includes the group size (adult and children), the stay length (count of weekend and weeknight stays), guest upgrades (parking, meal plans, room types and count of special requests), booking details (price and reservation method), the date of stay, and status (whether or not the guest cancelled the reservation).

Our team observed there was only one row of data for year 2015 and 2016. We additionally noticed there only 5 data points for meal plan 3. GroupSize, StayLength, HasChildren, HasRequests and Online were created from the existing dataset. GroupSize was an aggregation of count of adults and children, StayLength was an aggregation of Weekend and Weeknigths. HasChildren, HasRequests and Online created binary, true/false (0 and 1) variables from the count of children, requests and instances where the market was online.

The status column (cancelled verses not cancelled), is the response variable of interest in this analysis.

- *Goal of study*

1.1 Research Objectives

The goal of this analysis is to understand the best predictors for guest cancellations in order to develop company policy to decrease the cancellation rate. Lowering the percent cancellation rate to 20% represents a revenue increase of approximately $2 million dollars over four years.

What are the overarching research objectives that the study is targeting? (not just a restatement of the original research objectives)

Objective 1:

The first objective will explore and define the most important quantitative variable for predicting whether a customer will cancel their reservation.

Objective 2:

The second objective will explore and define the most important categorical variable for predicting whether a customer will cancel their reservation.

Objective 3:

The final goal of our statistical analysis will combine our understanding of quantitative and categorical variables in order to create a final statistical model to predict the customers who are most likely to cancel their reservations.

1.2 Variables

What is (are) possible explanatory and response variables?

The response variable of interest is status, whether or not a client cancelled their booking. The possible explanatory variables include both categorical and quantitative options. The number of requests from a client was converted to a categorical variable. A new column called HasRequests was created, with 1 indicating the customer had a special request. We made a similar to transformation to the children column, transforming this into a binary column indicating whether or not there were children in the group.

Create a table here - it should include variable names and types

	Description of Variable	Variable Type
Status	Cancellation status of booking	C
Meal	Type of meal in the booking	C
Parking	Whether or not a parking space is in the booking	C
RoomType	Type of room booked	C
Market	Market segment of booking	C
Month	Month of booking date	C
Requests	Num. of special requests made by the customer	C
HasChildren	Indicates whether or not the booking has children	C
HasRequests	Whether the booking had any special requests	C
Online	Indicator for an online booking	C
Adults	Number of adults in the booking	Q
Children	Number of children in the booking	Q
GroupSize	Number of people in the booking	Q

	Description of Variable	Variable Type
StayLength	Total of weekend plus weeknights	Q
Weekends	Number of weekend nights in the booking	Q
Weeknights	Number of weeknights in the booking	Q
LeadTime	Number of days between booking date and arrival	Q
AvgPrice	Average room price on the week of the booking	Q
Year	Year of the reservation	Q

2. Detailed Exploratory Data Analysis (EDA)

Spend time investigating the data thoroughly and use this section to present figures that allow you to visualize the relationship of the response to different predictors. Include descriptions of the figures that illustrate what you can learn from them.

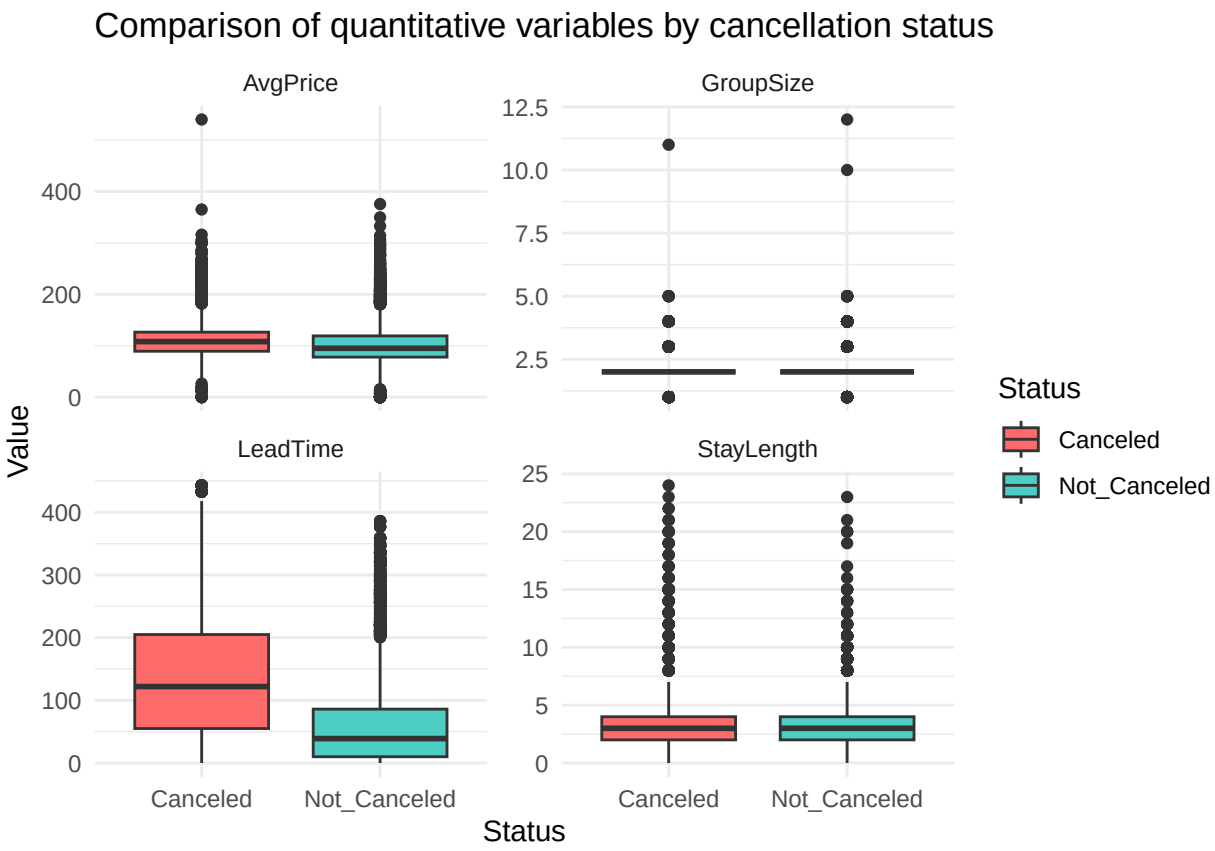
Quantitative EDA

For our quantitative exploration, we focused on the aggregated data columns of group size and stay length, in addition to the average price and lead time variables. We did not address children in the quantitative exploration, because we converted the presence of children into categorical variable of HasChildren.

Looking at the side by side box plot, the lead time has the most value for predicting cancellations.

Table 2: Quantitative Explanatory Variables

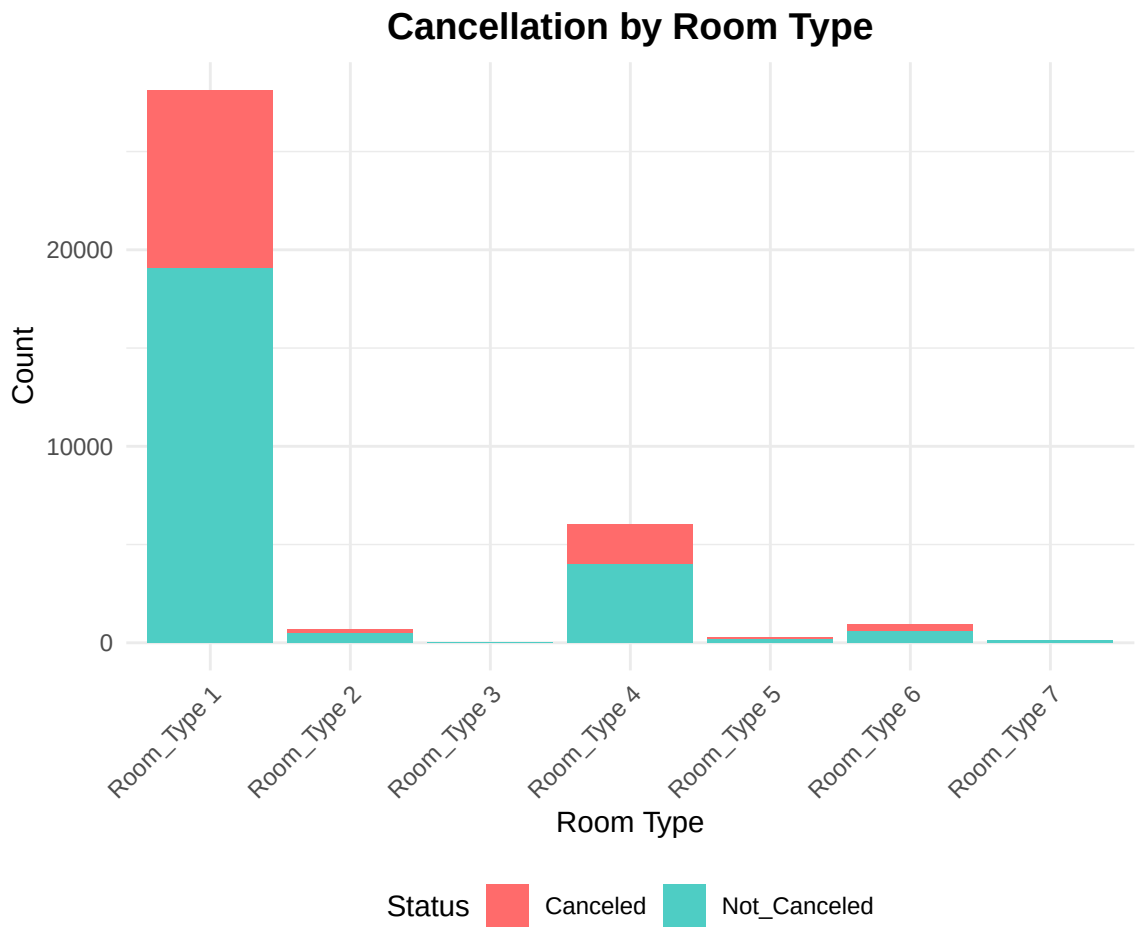
Status	Mean_LeadTime	Mean_AvgPrice	Mean_GroupSize	Mean_StayLength
Canceled	139.220	110.580	2.034	3.280
Not_Canceled	58.934	99.933	1.909	2.886

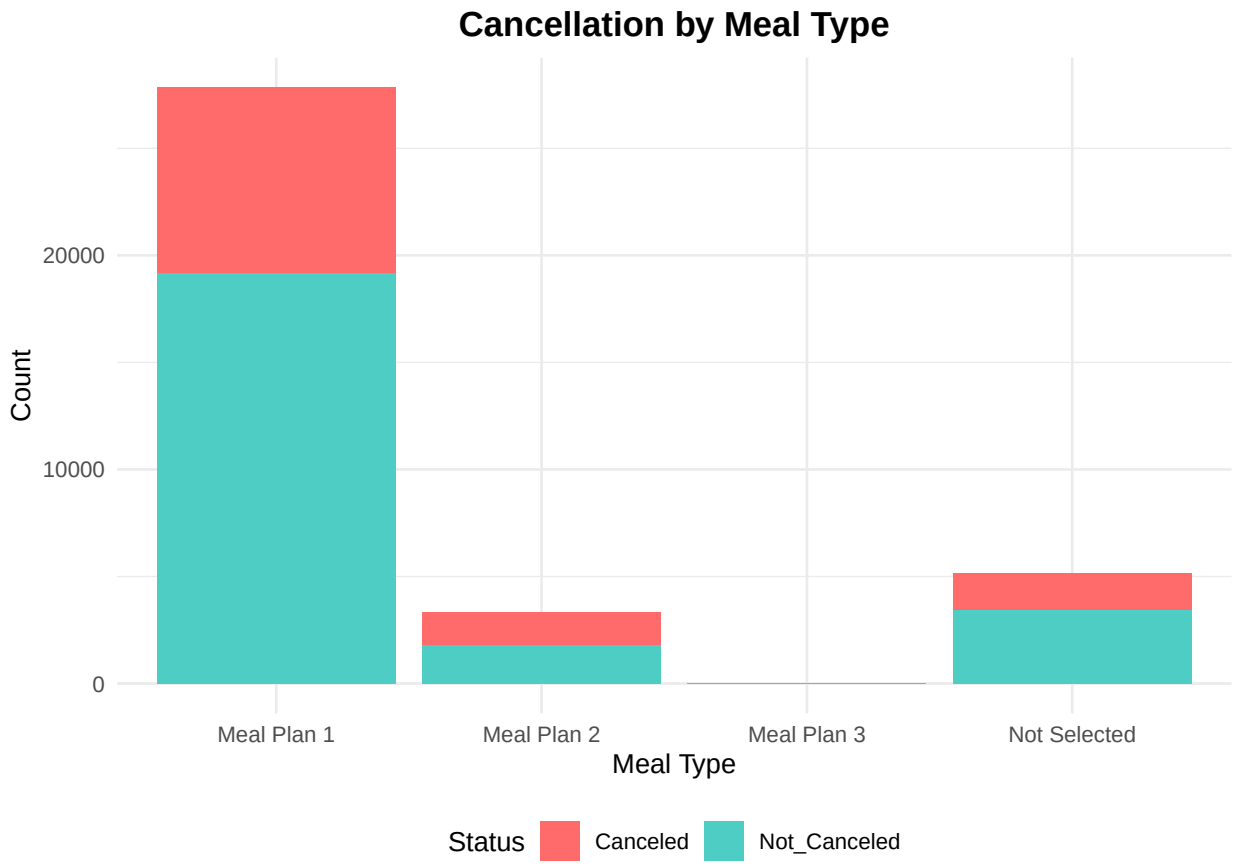


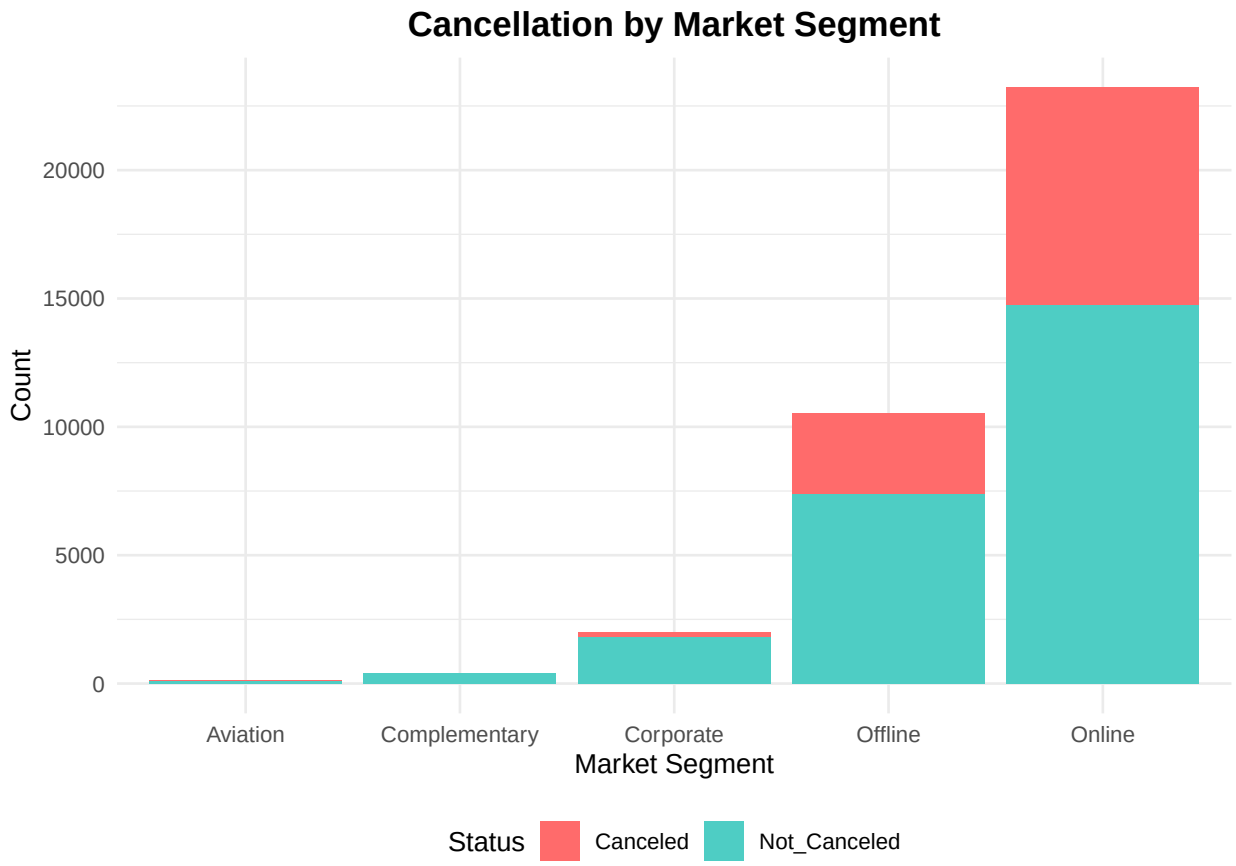
Categorical EDA

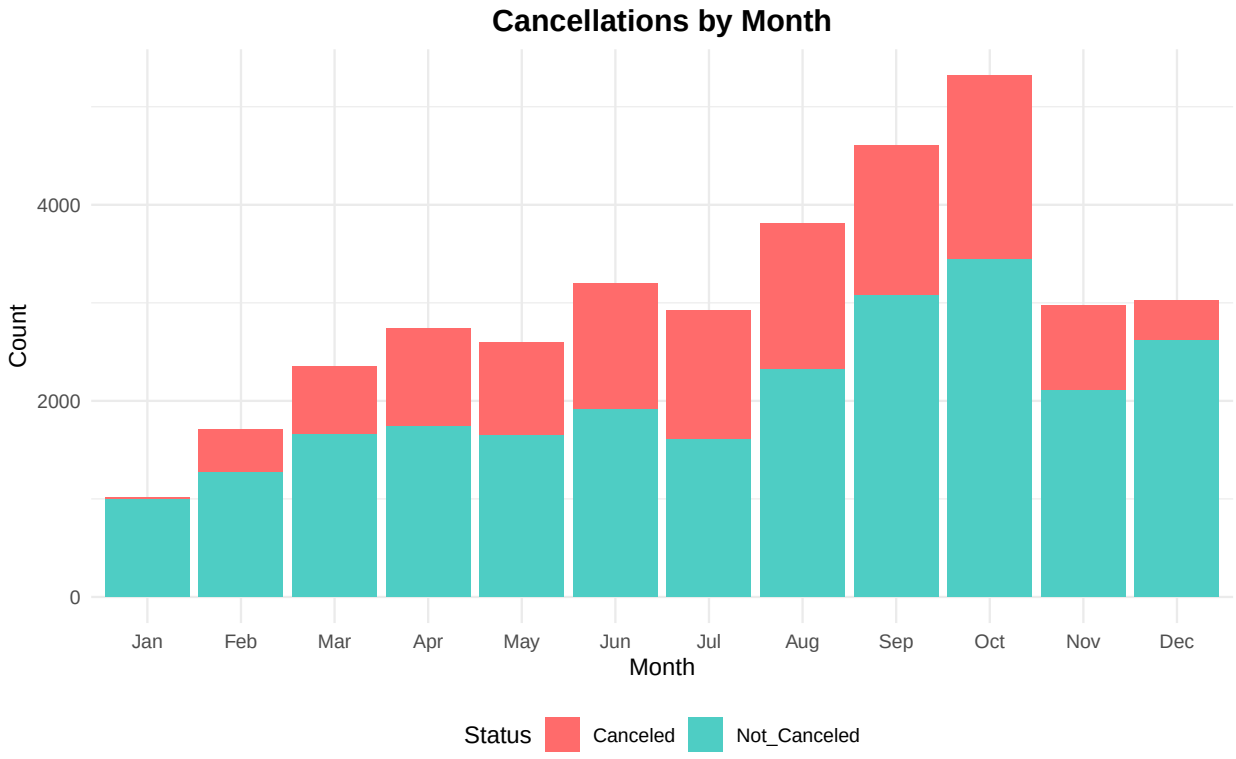
For our categorical exploration, we created side by side bar plots for room type, meal, market, month, request and if the group had children.

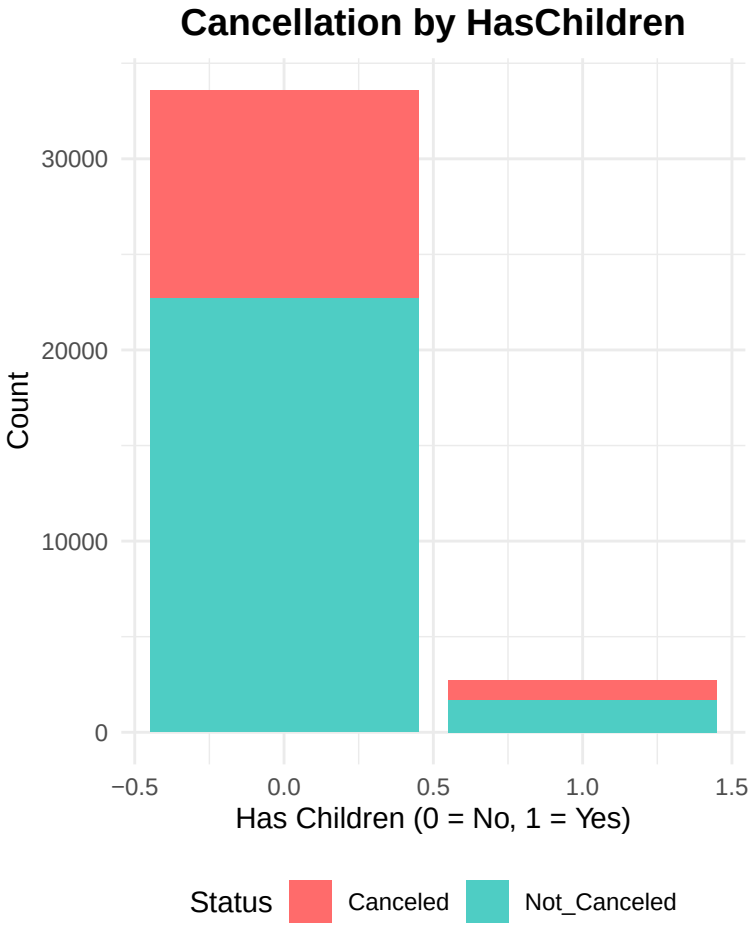
Looking at the side by side bar plots, the requests appeared to be the most useful variable, followed by month and market. After investigating the Online column created from Market, using the non-transformed Market column with all of the variables was more useful for our statistical model.

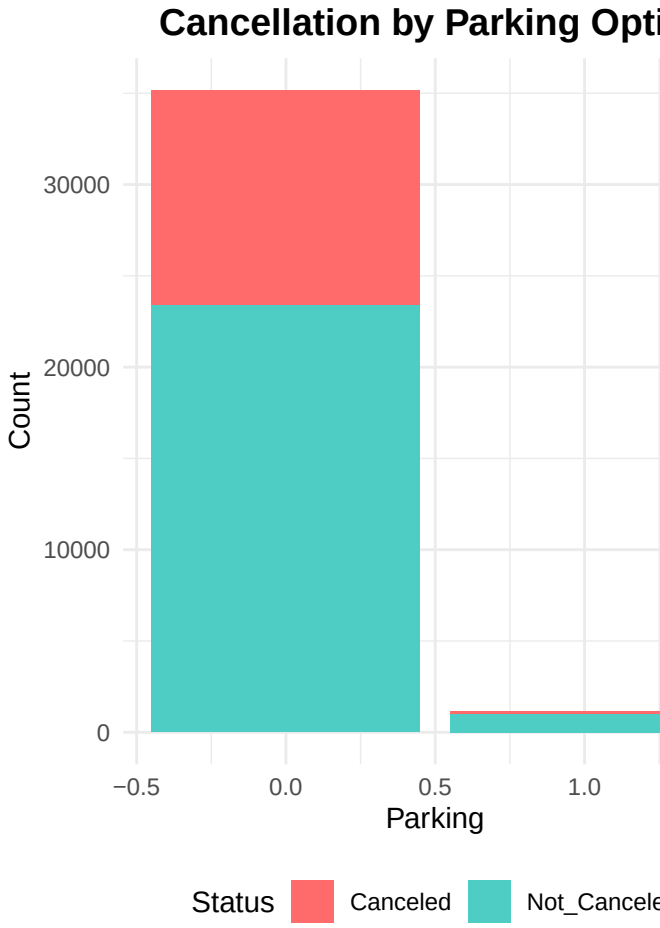

Cancellation by Room Type
Count
20000
10000
0
Room_Type 1
Room_Type 2
Room_Type 3
Room_Type 4
Room_Type 5
Room_Type 6
Room_Type 7
Room Type
Status
Canceled
Not_Canceled


Cancellation by Meal Type
Count
20000
10000
0
Meal Plan 1
Meal Plan 2
Meal Plan 3
Not Selected
Meal Type
Status
Canceled
Not_Canceled


Cancellation by Market Segment
20000
15000
10000
5000
0
Count
Aviation
Complementary
Corporate
Offline
Online
Market Segment
Status
Canceled
Not_Canceled


Cancellations by Month
4000
2000
0
Count
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Status
Canceled
Not_Canceled


Cancellation by HasChildren
30000
20000
10000
0
Count
−0.5
0.0
0.5
1.0
1.5
Has Children (0 = No, 1 = Yes)
Status
Canceled
Not_Canceled


Cancellation by Parking Opti
30000
20000
10000
0
Count
−0.5
0.0
0.5
1.0
Parking
Status
Canceled
Not_Cancele

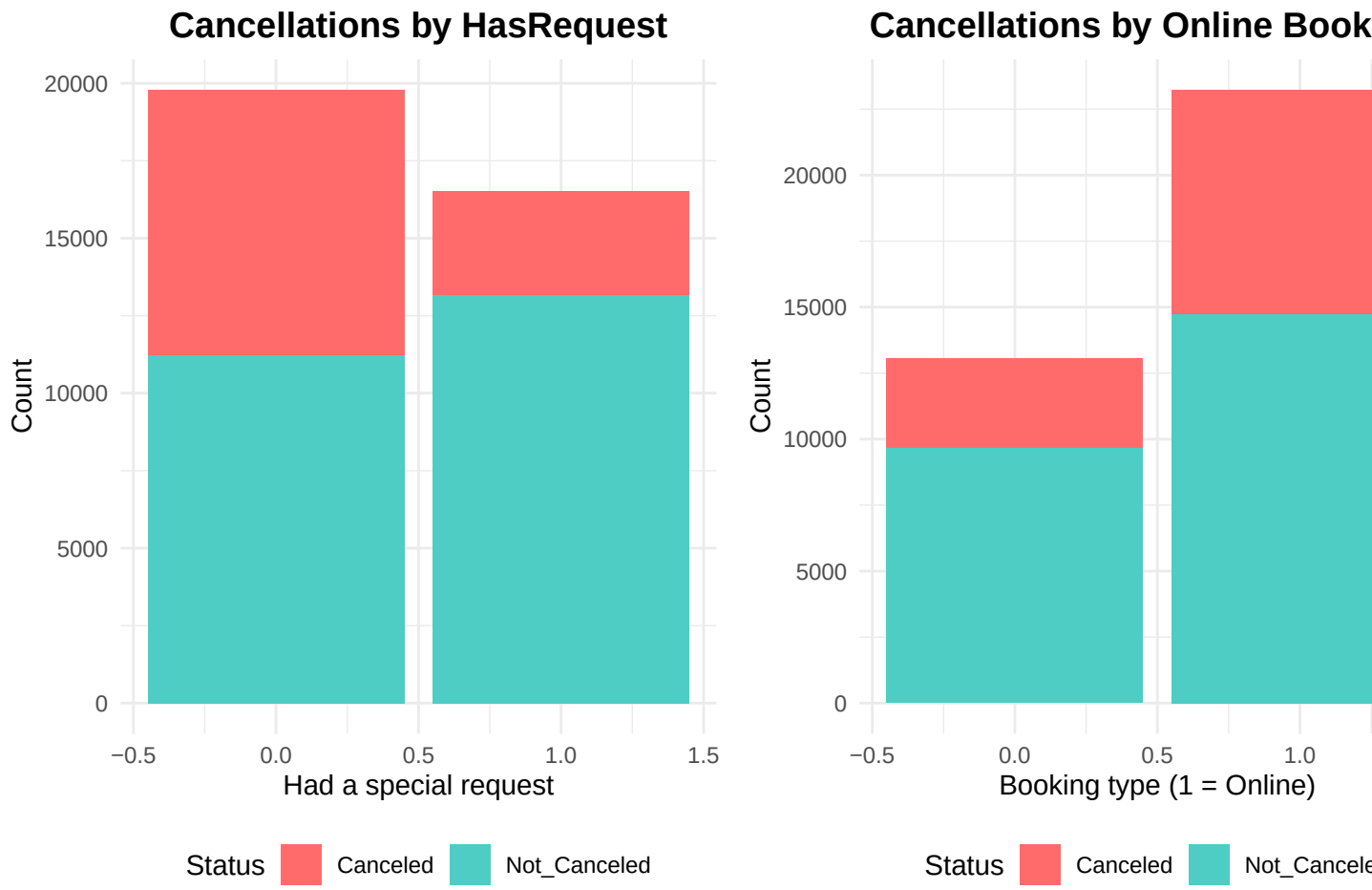


Table 3: Room Type

RoomType	Status	count	proportion
Room_Type 1	Canceled	9076	0.323
Room_Type 1	Not_Canceled	19062	0.677
Room_Type 2	Canceled	228	0.329
Room_Type 2	Not_Canceled	464	0.671
Room_Type 3	Canceled	2	0.286
Room_Type 3	Not_Canceled	5	0.714
Room_Type 4	Canceled	2069	0.341
Room_Type 4	Not_Canceled	3990	0.659
Room_Type 5	Canceled	72	0.272
Room_Type 5	Not_Canceled	193	0.728
Room_Type 6	Canceled	406	0.420
Room_Type 6	Not_Canceled	560	0.580
Room_Type 7	Canceled	36	0.228
Room_Type 7	Not_Canceled	122	0.772

Table 4: Meal Option

Meal	Status	count	proportion
Meal Plan 1	Canceled	8681	0.312
Meal Plan 1	Not_Canceled	19161	0.688
Meal Plan 2	Canceled	1507	0.456
Meal Plan 2	Not_Canceled	1799	0.544
Meal Plan 3	Canceled	1	0.200
Meal Plan 3	Not_Canceled	4	0.800
Not Selected	Canceled	1700	0.331
Not Selected	Not_Canceled	3432	0.669

Table 5: Market

Market	Status	count	proportion
Aviation	Canceled	37	0.296
Aviation	Not_Canceled	88	0.704
Complementary	Not_Canceled	391	1.000
Corporate	Canceled	220	0.109
Corporate	Not_Canceled	1797	0.891
Offline	Canceled	3154	0.299
Offline	Not_Canceled	7377	0.701
Online	Canceled	8478	0.365
Online	Not_Canceled	14743	0.635

Table 6: Parking

Parking	Status	count	proportion
0	Canceled	11775	0.335
0	Not_Canceled	23386	0.665
1	Canceled	114	0.101
1	Not_Canceled	1010	0.899

Table 7: Month

Month	Status	count	proportion
Jan	Canceled	24	0.024
Jan	Not_Canceled	990	0.976
Feb	Canceled	431	0.253
Feb	Not_Canceled	1274	0.747
Mar	Canceled	700	0.297
Mar	Not_Canceled	1658	0.703
Apr	Canceled	996	0.364
Apr	Not_Canceled	1741	0.636
May	Canceled	949	0.365
May	Not_Canceled	1650	0.635
Jun	Canceled	1291	0.403
Jun	Not_Canceled	1912	0.597
Jul	Canceled	1314	0.450

Jul	Not_Canceled	1607	0.550
Aug	Canceled	1488	0.390
Aug	Not_Canceled	2325	0.610
Sep	Canceled	1539	0.334
Sep	Not_Canceled	3073	0.666
Oct	Canceled	1880	0.353
Oct	Not_Canceled	3440	0.647
Nov	Canceled	875	0.294
Nov	Not_Canceled	2106	0.706
Dec	Canceled	402	0.133
Dec	Not_Canceled	2620	0.867

Table 8: Special Requests

HasRequests	Status	count	proportion
0	Canceled	8547	0.432
0	Not_Canceled	11233	0.568
1	Canceled	3342	0.202
1	Not_Canceled	13163	0.798

Table 9: Has Children

HasChildren	Status	count	proportion
0	Canceled	10885	0.324
0	Not_Canceled	22698	0.676
1	Canceled	1004	0.372
1	Not_Canceled	1698	0.628

Table 10: Online Booking

Online	Status	count	proportion
0	Canceled	3411	0.261
0	Not_Canceled	9653	0.739
1	Canceled	8478	0.365
1	Not_Canceled	14743	0.635

3. Statistical Analysis

Describe the statistical analysis that you used to answer Research Objectives 1-3. Each analysis should be labeled for the appropriate research objective. This section should summarize all relevant analyses that lead to your final conclusions/decisions/recommendations. Be sure to include:

- *Model Assumptions & how checked/verified*
- *Interpretation of relevant estimates/statistics/p-values IN CONTEXT*
- *Any data manipulation to create new variables*

You should not include your code. However, I would like to see the code for final models that you are using to answer the research objectives in the appendix.

3.1 Objective 1

The main values that we used to determine the best quantitative predictor were AIC and residual deviance; most of the models we created to find the best predictor of this nature yielded p-values that were lower than 2e-16. However, the AIC and residual deviance for the generalized linear model predicting Status with LeadTime was significantly lower than the values for AvgPrice, StayLength (the new variable synthesized from the number of weekdays and weeknights of each booking), and GroupSize (the new variable combining Adults and Children), reassuring LeadTime's strength as a predictor. With respect to the conditions of this model using only LeadTime, there is some concern about the equal variance condition; the values seem to be exhibiting some flaring near the right edge of the graph. We tried a square-root transformation on this model, but there was little to no improvement, so this was left out of the final model.

Logit function for Status using LeadTime:

$$\widehat{log(\frac{\pi}{1-\pi})} = 1.8046260 - 0.0117484(LeadTime)$$

$$odds(Status) = e^{1.8046260 - 0.0117484(LeadTime)}$$

Probability function for Status using LeadTime:

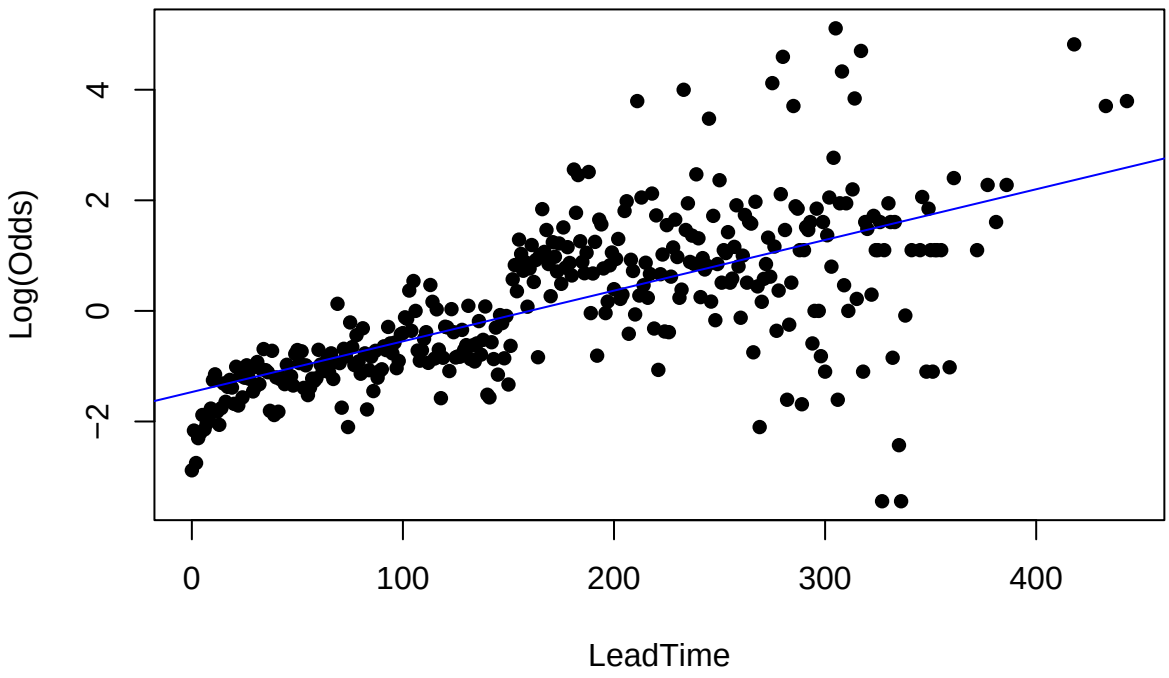
$$p(Status) = \widehat{\pi} = \frac{e^{1.8046260 - 0.0117484(LeadTime)}}{1 - e^{1.8046260 - 0.0117484(LeadTime)}}$$

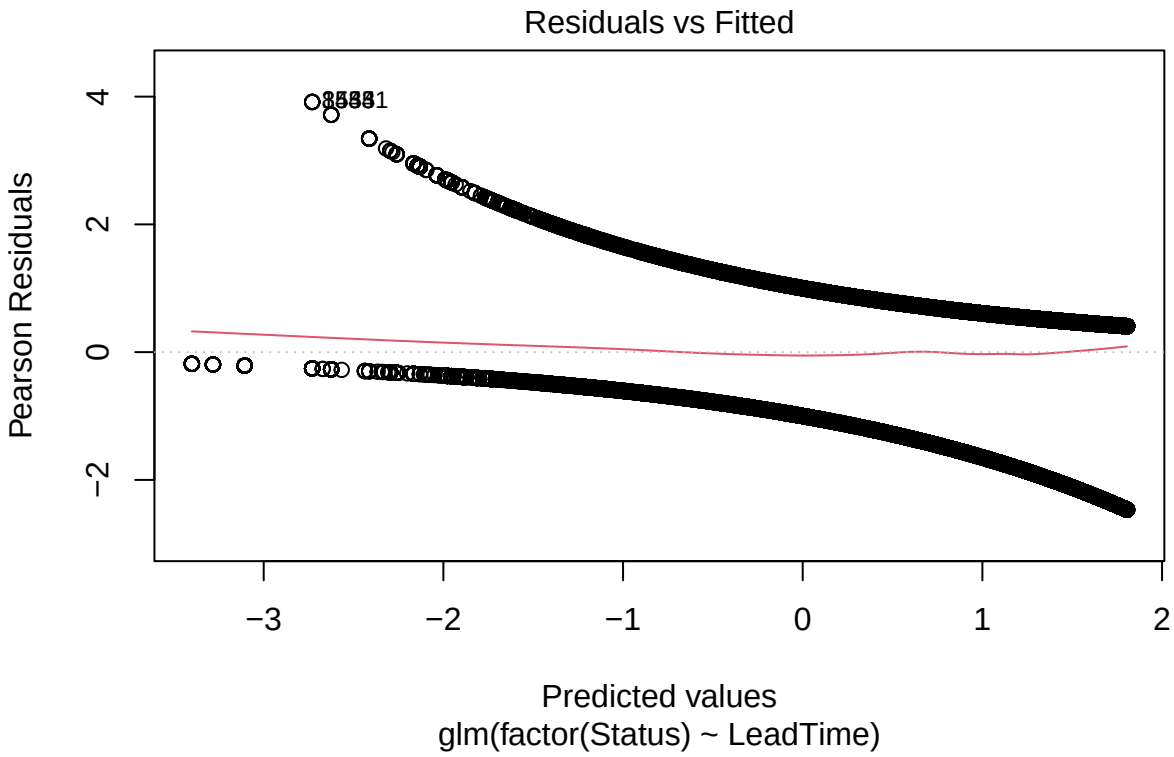
$$oddsratio = e^{\widehat{\beta_1}} = e^{-0.0117484} = 0.988324$$

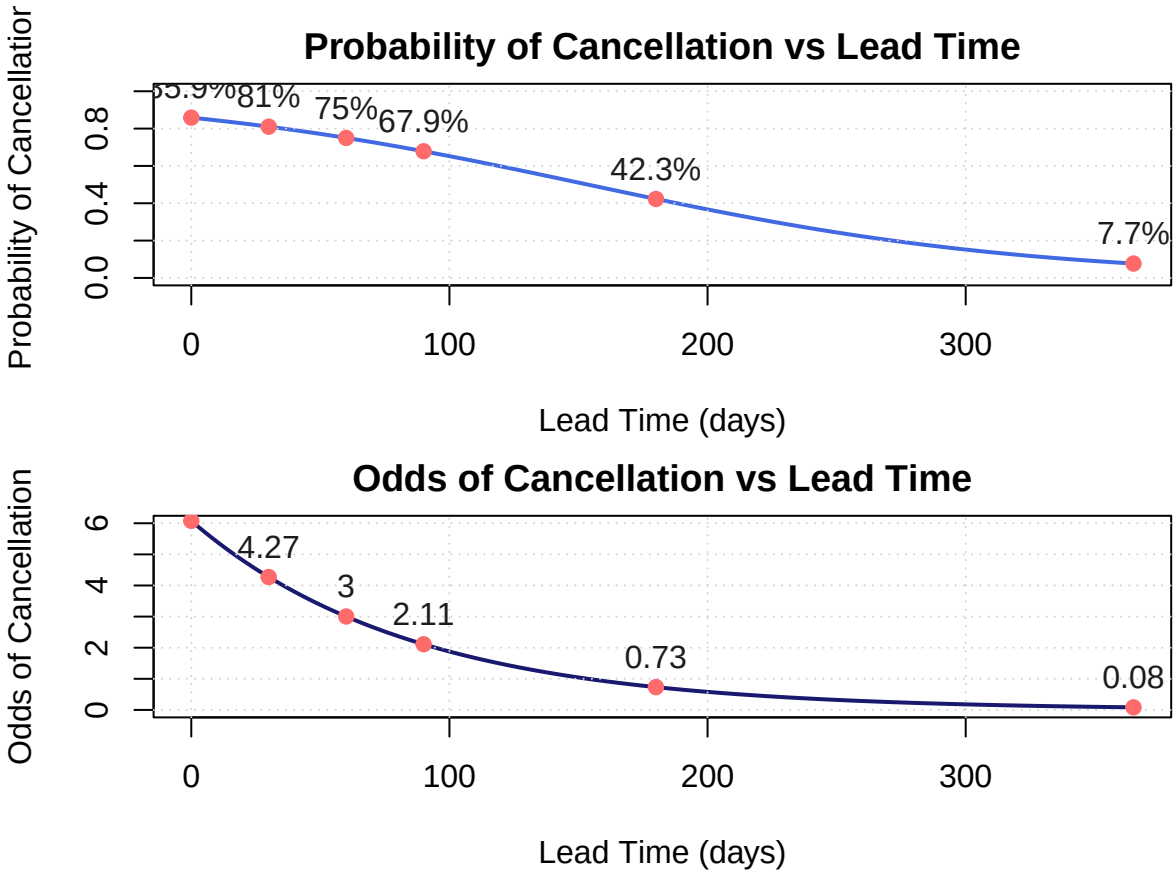
For every 1 day increase of LeadTime, the predicted odds of not cancelling the hotel booking decreases by a factor of 0.988. To further Gold Mine Resorts understanding of these statistics, Visuals of the probabilities and odds of cancellation are included below.

$$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$$

With a p-value $<$ 2eˆ-16, we have strong evidence to reject the null hypothesis in favor of the alternative. LeadTime is useful for predicting hotel cancellations.


Log(Odds)
−2
0
2
4
0
100
200
300
400
LeadTime


Residuals vs Fitted
Pearson Residuals
4
2
0
−2
−3
−2
−1
0
1
2
Predicted values
glm(factor(Status) ~ LeadTime)




```
##
## Percentage change in odds from Lead Time = 0:

## Lead Time = 30 days: -29.7% change in odds
## Lead Time = 60 days: -50.6% change in odds
## Lead Time = 90 days: -65.3% change in odds
## Lead Time = 180 days: -87.9% change in odds
## Lead Time = 365 days: -98.6% change in odds

##
## Odds ratio for one-day increase in lead time: 0.9883

## This corresponds to a 1.17% decrease in odds per day
```

3.2 Objective 2

Looking at the AIC for the categorical models, HasRequests was shown to be the best single categorical explanatory variable for predicting cancellations. As predicted by our exploratory data analysis, this was closely followed by Month and Market.

For our categorical analysis, we transformed Requests (a nominally quantitative variable) into a categorical variable, HasRequests, that indicates whether or not a given booking made a nonzero number of special requests. As evidenced by the model output for HasRequests, we note that the p-value for the variable's coefficient is virtually zero, meaning that HasRequests has an extremely statistically significant effect on

cancellation status. We are 95% confident that the aforementioned coefficient lies between 1.05 and 1.14, i.e. we are 95% certain that for bookings with special requests, the logarithmic likelihood of a cancellation will increase by a factor of between 1.05 and 1.14.

Logit function for Status using HasRequests:

$$\widehat{log(\frac{\pi}{1-\pi})} = 0.27328 + 1.09756(HasRequests)$$

$$odds(Status) = e^{0.27328+1.09756(HasRequests)}$$

Probability function for Status using HasRequests:

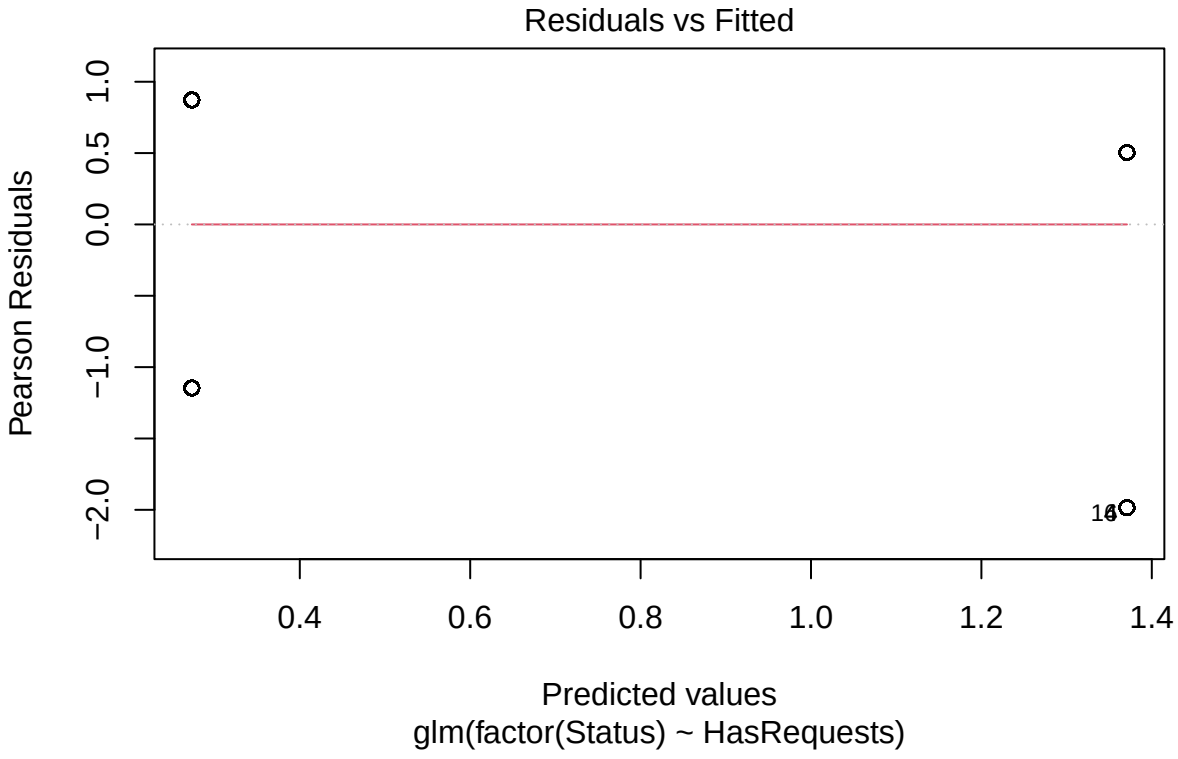
$$p(Status) = \widehat{\pi} = \frac{e^{0.27328+1.09756(HasRequests)}}{1 - e^{0.27328+1.09756(HasRequests)}}$$

$$oddsratio = e^{\widehat{\beta_1}} = e^{1.09756} = 2.996845$$

HasRequests is a binary categorical variable. When HasRequests = 1 the predicted odds of not cancelling are approximately 3 times higher than when HasRequests = 0.

$$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$$

With a p-value < 2eˆ-16, we have strong evidence to reject the null hypothesis in favor of the alternative. HasRequests is useful for predicting hotel cancellations.



3.3 Objective 3

When observing the ANOVA table, the HasChildren variable was not statistically significant. We replaced the HasChildren variable with Children and still saw observed a large p value, greater than .2 in both instances. We additionally removed year from the statistical model, since this historical data point is not useful for future predictions.

AIC with the forward step analysis recommended using all the remaining variables, including LeadTime, HasRequests, Market, Month, AvgPrice, Parking, RoomType, StayLength, Meal and GroupSize. The AIC of the model was 30,674.

The large number of variables made us concerned about the potential of overfitting. We used the output from an ANOVA table to eliminate variables one at a time, starting with GroupSize (the highest p value). The final model we selected used LeadTime, Market, Requests, Month, AvgPrice, and StayLength.

While our AIC increased slightly to 31,109, we reduced the variable count from 10 to 6. The 6 variable model had optimized p values for each explanatory variable. Additionally, our model had good variance and normality, in addition to fitting closely to the expected results (as indicated by the empirical logit plot).

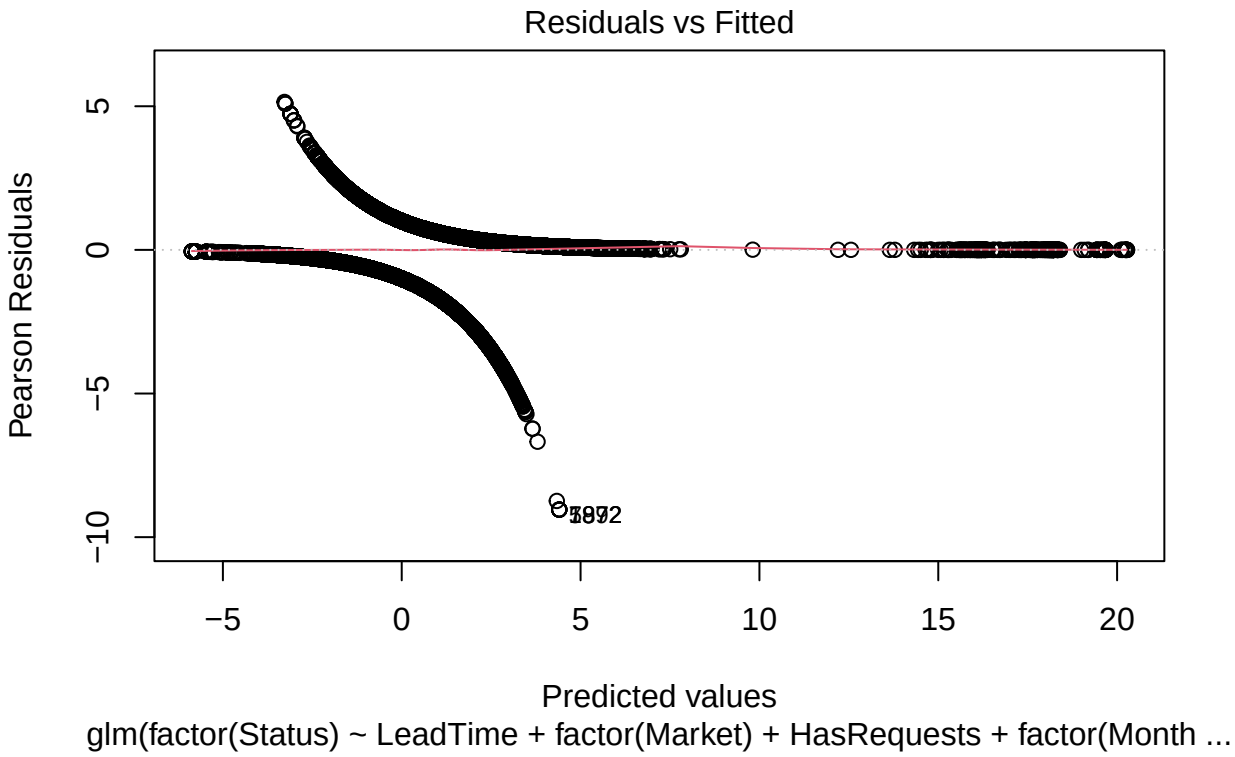
The data was divided into ten groups for cross-validation to test the predicted results compared to the actual and produced the following results:

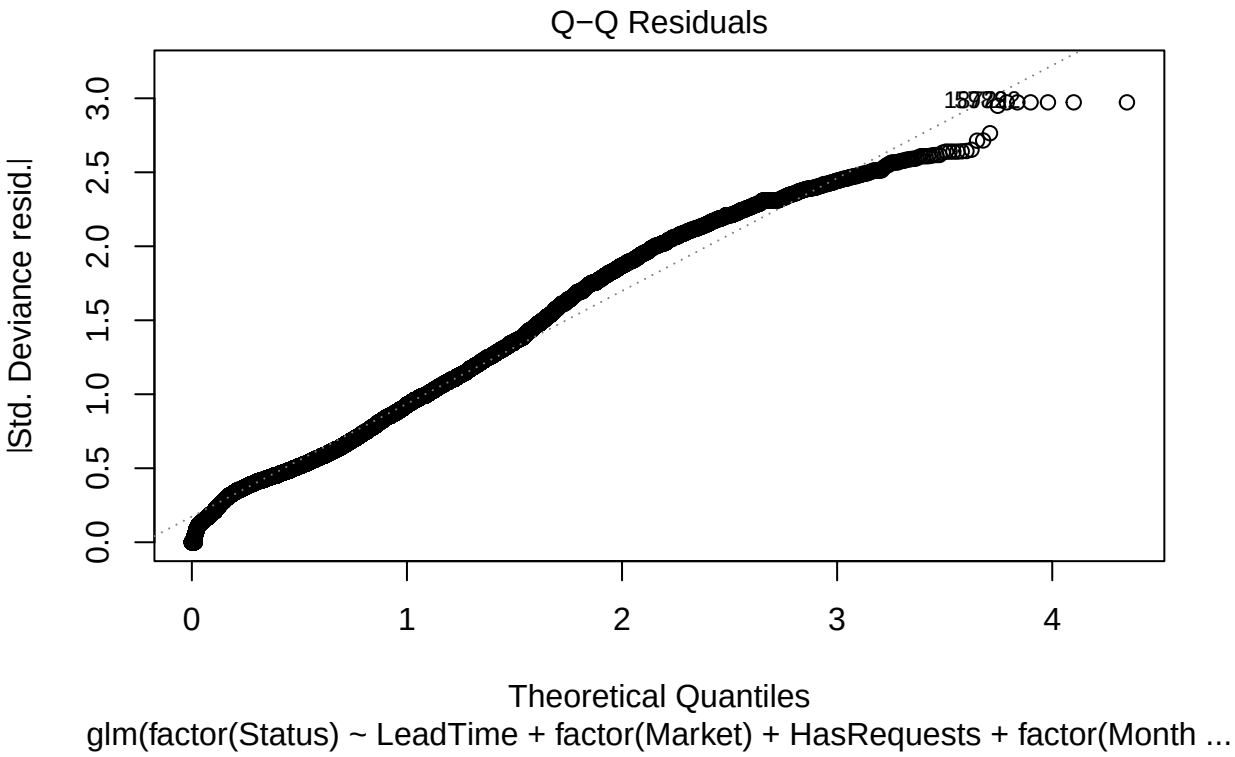
- Accuracy: 0.806
- Sensitivity: 0.894
- Specificity: 0.626
- AUC: 0.859

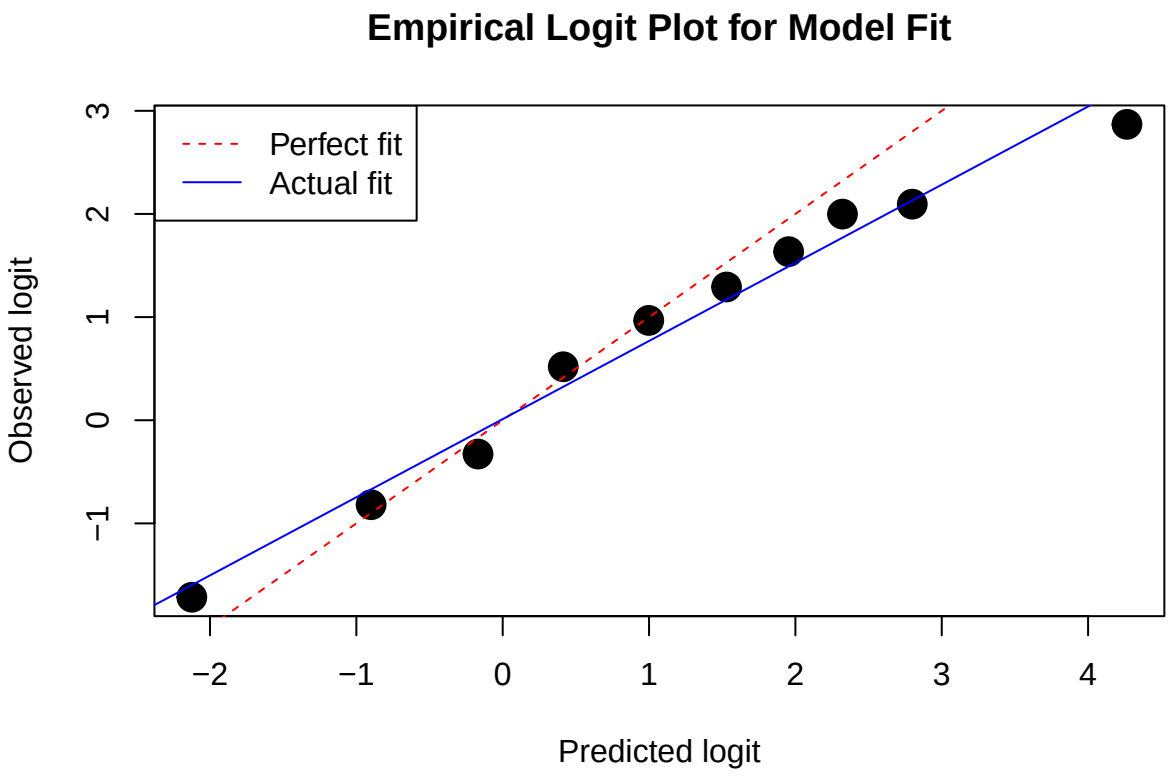
The high sensitivity indicates that the model is very good at identifying reservations that will not be cancelled. The specificity value of .626 means our model is more likely to falsely predict a customer will not cancel when they end up cancelling.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: factor(Status)
##                 LR Chisq Df Pr(>Chisq)
## LeadTime         7937.8  1  < 2.2e-16 ***
## factor(Market)   2553.0  4  < 2.2e-16 ***
## HasRequests      4287.6  1  < 2.2e-16 ***
## factor(Month)    1097.4 11  < 2.2e-16 ***
## AvgPrice          945.0  1  < 2.2e-16 ***
## StayLength         65.1  1  7.202e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 31109.38
```


Residuals vs Fitted
Pearson Residuals
5
0
−5
−10
−5
0
5
10
15
20
Predicted values
glm(factor(Status) ~ LeadTime + factor(Market) + HasRequests + factor(Month ...


Q−Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
Theoretical Quantiles
glm(factor(Status) ~ LeadTime + factor(Market) + HasRequests + factor(Month ...




```
## Warning: package 'pROC' was built under R version 4.4.3

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:mosaic':
##
##     cov, var

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## [1] "Cross-validation results:"

##    fold  accuracy sensitivity specificity       auc
## 1     1 0.8076379   0.8941469   0.6225577 0.8615138
## 2     2 0.8034072   0.8960776   0.6163265 0.8533144
## 3     3 0.7960077   0.8869919   0.6090226 0.8505233
## 4     4 0.8138298   0.8983891   0.6359687 0.8634720
## 5     5 0.8121730   0.9011676   0.6390916 0.8681748
## 6     6 0.8114047   0.9033207   0.6315789 0.8664678
## 7     7 0.8000000   0.8720456   0.6521739 0.8568420
## 8     8 0.8088719   0.8949479   0.6234888 0.8581146
## 9     9 0.7950549   0.8899044   0.6098946 0.8485808
## 10   10 0.8120383   0.8988405   0.6239168 0.8602472

## [1] "Average performance:"

##    accuracy sensitivity specificity         auc
##   0.8060425   0.8935832   0.6264020   0.8587251
```

```
## [1] "Standard deviation of performance:"
```

Key Quantitative Predictors:*

$$oddsratio_{LeadTime} = e^{-0.0161356} = 0.984$$

For each additional day in lead time, the predicted odds of not cancelling decrease by a factor of 0.984, holding all other variables constant.

$$oddsratio_{AvgPrice} = e^{-0.0145549} = 0.985$$

For each $1 increase in average price, the predicted odds of not cancelling decrease by a factor of 0.985, holding all other variables constant.

$$oddsratio_{StayTime} = e^{-0.0643062} = 0.938$$

For each additional night in the hotel, the predicted odds of not cancelling decrease by a factor of 0.938, holding all other variables constant.

To summarize, the business implication is that longer lead times, higher prices, and longer stays correlate with higher cancellation risk. In addition, LeadTime, AvgPrice, and StayTime have respective p-values of <2e-16, <2e-16, and 7.49e-16, indicating that these variables are useful to predict Status.

Key Categorical Predictors:

$$oddsratio_{HasRequests} = e^{2.0247462} = 7.57$$

Holding all other variables constant, when a guest makes at least one special request the predicted odds of not cancelling are approximately 8 times higher than those without special requests. With a p-value <2e-16, this variable is useful to predict Status.

$$oddsratio_{Market(Corporate)} = e^{1.0866028} = 2.96$$

Holding all other variables constant, when Gold Mine Resorts has a corporate booking the predicted odds of not cancelling are approximately 3 times higher than an aviation booking (baseline). With a p-value of 3.95e-07, this segment is useful in predicting Status.

$$oddsratio_{Market(Offline)} = e^{1.6324226} = 5.12$$

Holding all other variables constant, when Gold Mine Resorts has an offline booking the predicted odds of not cancelling are approximately 5 times higher than an aviation booking (baseline). With a p-value of 8.80e-16, this segment is useful in predicting Status.

$$oddsratio_{Market(Online)} = e^{-0.2199951} = 0.8$$

Holding all other variables constant, when Gold Mine Resorts has an online booking the predicted odds of not cancelling decrease by a factor of 0.8 compared to an aviation booking (baseline). However, with a p-value of 0.27308, this segment is not statistically significant in predicting Status.

$$oddsratio_{Month(Feb-Nov)} \approx e^{-2.28} = 0.102$$

Holding all other variables constant, when there is a booking made between February to November (all β's are relatively similar in range) the predicted odds of not cancelling decrease by a factor of 0.102 when compared to bookings made in January (baseline). Each individual month is statistically significant in predicting Status with p-values <2e-16.

$$oddsratio_{Month(Dec)} \approx e^{-0.0643} = 0.55$$

Holding all other variables constant, when there is a booking made in December the predicted odds of not cancelling decrease by a factor of 0.55 when compared to bookings made in January (baseline). December is a statistically significant segment of Month in predicting Status with a p-value of 0.00674.

$$H_0 : \beta_k = 0, H_a : \beta_k \neq 0$$

To summarize, the business implication is that corporate/offline bookings are more reliable than aviation bookings. In addition, the months of February to November have a higher cancellation risk than January. Specifically, the peak cancellation risk occurs in the month of February where the predicted odds of not cancelling decrease by a factor of 0.059 compared to January (holding all other variables constant).

4. Conclusions

Succinct response to each question laid out in 1.1. This a much shorter version of section 3, and focuses on conclusions rather than the analysis.

Objective 1:

Our final conclusion about Objective 1 is that for the reasons described earlier, LeadTime is the best single quantitative predictor of whether or not a given booking will be cancelled, even when considering new variables synthesized from the dataset for the purpose of analysis.

```
##                   2.5 %      97.5 %
## (Intercept)  1.76657831  1.84267367
## LeadTime    -0.01206072 -0.01143611
```

```
##                 2.5 %   97.5 %
## (Intercept) 5.8507994 6.313396
## LeadTime    0.9880117 0.988629
```

Objective 2:

When considering only the categorical variables available to us from the dataset (and those that we synthesized ourselves), we found that HasRequests (i.e. whether a reservation has any special requests) is the most useful variable for predicting cancellation of bookings.

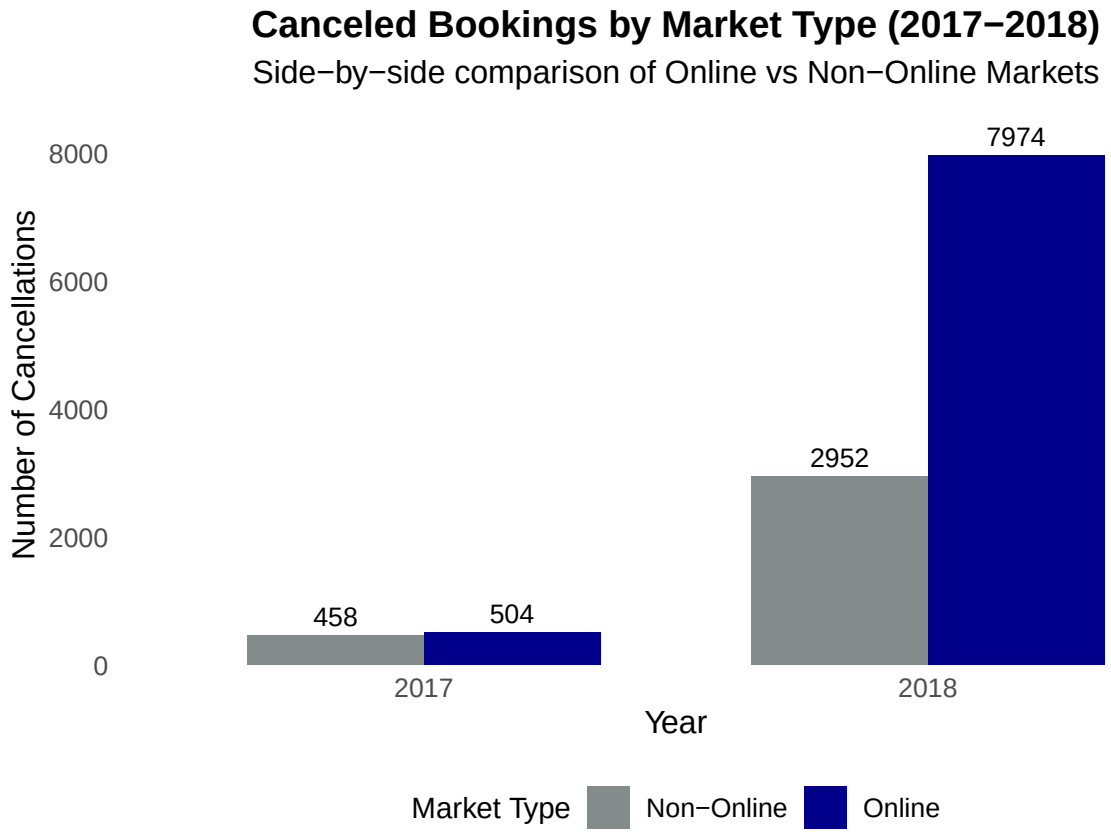
```
##                 2.5 %    97.5 %
## (Intercept) 0.2451431 0.3014079
## HasRequests 1.0503133 1.1448165
```

Objective 3:

With all of the original and synthesized variables available to us to pick for our final model, we observed above that the most useful model for predicting booking cancellation status includes the predictors of LeadTime, Market, Month, AvgPrice, and StayLength. Gold Mine Resorts should focus on these categories and note

that they are the keys to minimizing cancellations in the future. There additional factors, such as year, that have a role in whether or not people cancel, but these are not able to be controlled by Gold Mine Resorts.

While this model is a good starting point for predicting hotel cancellations, we noticed that there was a significant increase in both the count of reservations and the count of online reservations that led to cancellations from 2017 to 2018. If Gold Mine Resorts has access to further details about the sources for online bookings, this may prove to be a beneficial research study in the future.



5. Appendix

This section should include the code and output for the models used to answer the research objectives.

```
##
## Call:
## glm(formula = factor(Status) ~ AvgPrice, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.6315518  0.0365165   44.68   <2e-16 ***
## AvgPrice    -0.0086872  0.0003261  -26.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 45167  on 36283  degrees of freedom
## AIC: 45171
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = factor(Status) ~ LeadTime, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.8046260  0.0194124   92.96   <2e-16 ***
## LeadTime    -0.0117484  0.0001593  -73.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 38894  on 36283  degrees of freedom
## AIC: 38898
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = factor(Status) ~ StayLength, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.086290   0.022432   48.43   <2e-16 ***
## StayLength  -0.119786   0.006249  -19.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 45524  on 36283  degrees of freedom
## AIC: 45528
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = factor(Status) ~ GroupSize, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.29287    0.03602   35.89   <2e-16 ***
## GroupSize   -0.29141    0.01721  -16.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 45612  on 36283  degrees of freedom
## AIC: 45616
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = factor(Status) ~ Requests, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.27765    0.01392   19.95   <2e-16 ***
## Requests     0.84692    0.01826   46.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 43272  on 36283  degrees of freedom
## AIC: 43276
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = factor(Status) ~ RoomType, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.74206    0.01275  58.187  < 2e-16 ***
## RoomTypeRoom_Type 2 -0.03152    0.08188  -0.385  0.70022
## RoomTypeRoom_Type 3  0.17423    0.83676   0.208  0.83506
## RoomTypeRoom_Type 4 -0.08534    0.02994  -2.850  0.00437 **
## RoomTypeRoom_Type 5  0.24396    0.13868   1.759  0.07856 .
## RoomTypeRoom_Type 6 -0.42048    0.06642  -6.331 2.44e-10 ***
## RoomTypeRoom_Type 7  0.47844    0.19010   2.517  0.01184 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
```

```
## Residual deviance: 45845  on 36278  degrees of freedom
## AIC: 45859
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = factor(Status) ~ Parking, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.68616    0.01130   60.72   <2e-16 ***
## Parking      1.49535    0.09944   15.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 45574  on 36283  degrees of freedom
## AIC: 45578
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = factor(Status) ~ Market, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.86642    0.19594   4.422 9.78e-06 ***
## MarketComplementary 14.69965   73.60288   0.200    0.842
## MarketCorporate      1.23383    0.20855   5.916 3.29e-09 ***
## MarketOffline       -0.01672    0.19709  -0.085    0.932
## MarketOnline        -0.31313    0.19641  -1.594    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 44879  on 36280  degrees of freedom
## AIC: 44889
##
## Number of Fisher Scoring iterations: 14

##
## Call:
## glm(formula = factor(Status) ~ Meal, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
```

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.79174    0.01294  61.197  < 2e-16 ***
## MealMeal Plan 2  -0.61463    0.03724 -16.505  < 2e-16 ***
## MealMeal Plan 3   0.59455    1.11811   0.532  0.59490
## MealNot Selected -0.08923    0.03236  -2.758  0.00582 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 45634  on 36281  degrees of freedom
## AIC: 45642
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = factor(Status) ~ Month, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.7197     0.2066  18.006   <2e-16 ***
## MonthFeb     -2.6358     0.2140 -12.319   <2e-16 ***
## MonthMar     -2.8574     0.2114 -13.514   <2e-16 ***
## MonthApr     -3.1612     0.2104 -15.027   <2e-16 ***
## MonthMay     -3.1665     0.2106 -15.039   <2e-16 ***
## MonthJun     -3.3269     0.2097 -15.865   <2e-16 ***
## MonthJul     -3.5184     0.2099 -16.762   <2e-16 ***
## MonthAug     -3.2734     0.2092 -15.645   <2e-16 ***
## MonthSep     -3.0281     0.2089 -14.494   <2e-16 ***
## MonthOct     -3.1155     0.2086 -14.938   <2e-16 ***
## MonthNov     -2.8413     0.2105 -13.501   <2e-16 ***
## MonthDec     -1.8452     0.2134  -8.646   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45901  on 36284  degrees of freedom
## Residual deviance: 44226  on 36273  degrees of freedom
## AIC: 44250
##
## Number of Fisher Scoring iterations: 6

## Start:  AIC=45902.98
## factor(Status) ~ 1
##
##                    Df Deviance   AIC
## + LeadTime          1    38894 38898
## + HasRequests       1    43686 43690
## + factor(Month)    11    44226 44250
```

```
## + factor(Market)    4    44879 44889
## + AvgPrice          1    45167 45171
## + StayLength        1    45524 45528
## + factor(Parking)   1    45574 45578
## + GroupSize         1    45612 45616
## + factor(Meal)      3    45634 45642
## + factor(RoomType)  6    45845 45859
## <none>                   45901 45903
##
## Step:  AIC=38898.17
## factor(Status) ~ LeadTime
##
##                     Df Deviance   AIC
## + HasRequests        1    37078 37084
## + factor(Market)     4    37267 37279
## + AvgPrice           1    37491 37497
## + factor(Month)     11    37671 37697
## + factor(RoomType)   6    38630 38646
## + GroupSize          1    38661 38667
## + factor(Parking)    1    38697 38703
## + factor(Meal)       3    38750 38760
## + StayLength         1    38812 38818
## <none>                    38894 38898
##
## Step:  AIC=37084.44
## factor(Status) ~ LeadTime + HasRequests
##
##                     Df Deviance   AIC
## + factor(Market)     4    33257 33271
## + AvgPrice           1    35004 35012
## + factor(Month)     11    35794 35822
## + GroupSize          1    36436 36444
## + factor(RoomType)   6    36591 36609
## + factor(Meal)       3    36829 36841
## + StayLength         1    36909 36917
## + factor(Parking)    1    36970 36978
## <none>                    37078 37084
##
## Step:  AIC=33271.15
## factor(Status) ~ LeadTime + HasRequests + factor(Market)
##
##                     Df Deviance   AIC
## + factor(Month)     11    32063 32099
## + AvgPrice           1    32215 32231
## + factor(Parking)    1    33083 33099
## + GroupSize          1    33119 33135
## + factor(RoomType)   6    33133 33159
## + factor(Meal)       3    33162 33182
## + StayLength         1    33230 33246
## <none>                    33257 33271
##
## Step:  AIC=32098.81
## factor(Status) ~ LeadTime + HasRequests + factor(Market) + factor(Month)
##
```

```
##                     Df Deviance   AIC
## + AvgPrice           1    31134 31172
## + factor(Parking)    1    31885 31923
## + GroupSize          1    31899 31937
## + factor(RoomType)   6    31949 31997
## + factor(Meal)       3    31967 32009
## + StayLength         1    32014 32052
## <none>                    32063 32099
##
## Step:  AIC=31172.45
## factor(Status) ~ LeadTime + HasRequests + factor(Market) + factor(Month) +
##     AvgPrice
##
##                     Df Deviance   AIC
## + factor(Parking)    1    30890 30930
## + factor(RoomType)   6    31024 31074
## + factor(Meal)       3    31043 31087
## + StayLength         1    31069 31109
## <none>                    31135 31173
## + GroupSize          1    31134 31174
##
## Step:  AIC=30929.5
## factor(Status) ~ LeadTime + HasRequests + factor(Market) + factor(Month) +
##     AvgPrice + factor(Parking)
##
##                     Df Deviance   AIC
## + factor(RoomType)   6    30773 30825
## + factor(Meal)       3    30800 30846
## + StayLength         1    30834 30876
## <none>                    30890 30930
## + GroupSize          1    30889 30931
##
## Step:  AIC=30824.65
## factor(Status) ~ LeadTime + HasRequests + factor(Market) + factor(Month) +
##     AvgPrice + factor(Parking) + factor(RoomType)
##
##                 Df Deviance   AIC
## + StayLength     1    30696 30750
## + factor(Meal)   3    30712 30770
## + GroupSize      1    30764 30818
## <none>                30773 30825
##
## Step:  AIC=30749.68
## factor(Status) ~ LeadTime + HasRequests + factor(Market) + factor(Month) +
##     AvgPrice + factor(Parking) + factor(RoomType) + StayLength
##
##                 Df Deviance   AIC
## + factor(Meal)   3    30619 30679
## + GroupSize      1    30689 30745
## <none>                30696 30750
##
## Step:  AIC=30679.42
## factor(Status) ~ LeadTime + HasRequests + factor(Market) + factor(Month) +
##     AvgPrice + factor(Parking) + factor(RoomType) + StayLength +
```

```
##     factor(Meal)
##
##             Df Deviance   AIC
## + GroupSize  1    30613 30675
## <none>            30619 30679
##
## Step:  AIC=30674.51
## factor(Status) ~ LeadTime + HasRequests + factor(Market) + factor(Month) +
##     AvgPrice + factor(Parking) + factor(RoomType) + StayLength +
##     factor(Meal) + GroupSize


##
## Call:  glm(formula = factor(Status) ~ LeadTime + HasRequests + factor(Market) +
##     factor(Month) + AvgPrice + factor(Parking) + factor(RoomType) +
##     StayLength + factor(Meal) + GroupSize, family = binomial(link = "logit"),
##     data = hotel_bookings)
##
## Coefficients:
##                 (Intercept)                     LeadTime
##                     5.15790                     -0.01651
##                HasRequests  factor(Market)Complementary
##                     2.02060                     17.07502
##     factor(Market)Corporate        factor(Market)Offline
##                     1.14786                      1.91428
##        factor(Market)Online             factor(Month)Feb
##                     0.06250                     -2.83161
##            factor(Month)Mar             factor(Month)Apr
##                    -2.52033                     -2.28402
##            factor(Month)May             factor(Month)Jun
##                    -1.90397                     -2.13202
##            factor(Month)Jul             factor(Month)Aug
##                    -1.98030                     -1.92694
##            factor(Month)Sep             factor(Month)Oct
##                    -1.73622                     -1.97145
##            factor(Month)Nov             factor(Month)Dec
##                    -2.39601                     -0.50040
##                    AvgPrice             factor(Parking)1
##                    -0.01911                      1.58065
## factor(RoomType)Room_Type 2  factor(RoomType)Room_Type 3
##                     0.28519                      0.17572
## factor(RoomType)Room_Type 4  factor(RoomType)Room_Type 5
##                     0.25314                      0.75040
## factor(RoomType)Room_Type 6  factor(RoomType)Room_Type 7
##                     0.85804                      1.45018
##                   StayLength      factor(Meal)Meal Plan 2
##                     -0.07773                     -0.15018
##     factor(Meal)Meal Plan 3     factor(Meal)Not Selected
##                   -10.60678                     -0.36584
##                   GroupSize
##                    -0.07558
##
## Degrees of Freedom: 36284 Total (i.e. Null);  36254 Residual
## Null Deviance:       45900
## Residual Deviance: 30610     AIC: 30670
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: factor(Status)
##                   LR Chisq Df Pr(>Chisq)
## factor(Meal)         76.1  3  < 2.2e-16 ***
## factor(Parking)     238.4  1  < 2.2e-16 ***
## factor(RoomType)    109.8  6  < 2.2e-16 ***
## LeadTime           7727.5  1  < 2.2e-16 ***
## factor(Market)     1958.1  4  < 2.2e-16 ***
## AvgPrice            919.6  1  < 2.2e-16 ***
## HasRequests        4127.0  1  < 2.2e-16 ***
## factor(Month)      1109.9 11  < 2.2e-16 ***
## StayLength           90.7  1  < 2.2e-16 ***
## GroupSize             6.9  1   0.008542 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table (Type II tests)
##
## Response: factor(Status)
##                 LR Chisq Df Pr(>Chisq)
## LeadTime         7937.8  1  < 2.2e-16 ***
## factor(Market)   2553.0  4  < 2.2e-16 ***
## HasRequests      4287.6  1  < 2.2e-16 ***
## factor(Month)    1097.4 11  < 2.2e-16 ***
## AvgPrice          945.0  1  < 2.2e-16 ***
## StayLength         65.1  1  7.202e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 31109.38
```