Homework 3 – Sample Solution 1

Due Thursday, May 28th at 11:59 PM

This homework is supposed to be more comprehensive in nature, covering Lectures 2-16 and Labs 2-8. Please write your solutions in an R Markdown file and submit your PDF output to Gradescope. All code for your solutions should be shown in your PDF file.

```
## Rows: 119,390
## Columns: 32
## $ hotel                           <chr> "Resort Hotel", "Resort Hotel", "Resort~
## $ is_canceled                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, ~
## $ lead_time                       <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, ~
## $ arrival_date_year               <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 201~
## $ arrival_date_month              <chr> "July", "July", "July", "July", "July",~
## $ arrival_date_week_number        <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,~
## $ arrival_date_day_of_month       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ stays_in_weekend_nights         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ stays_in_week_nights            <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, ~
## $ adults                          <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ children                        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ babies                          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ meal                            <chr> "BB", "BB", "BB", "BB", "BB", "BB", "BB~
## $ country                         <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GBR~
## $ market_segment                  <chr> "Direct", "Direct", "Direct", "Corporat~
## $ distribution_channel            <chr> "Direct", "Direct", "Direct", "Corporat~
## $ is_repeated_guest               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_cancellations          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_bookings_not_canceled <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ reserved_room_type              <chr> "C", "C", "A", "A", "A", "A", "C", "C",~
## $ assigned_room_type              <chr> "C", "C", "C", "A", "A", "A", "C", "C",~
## $ booking_changes                 <dbl> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ deposit_type                    <chr> "No Deposit", "No Deposit", "No Deposit~
## $ agent                           <chr> "NULL", "NULL", "NULL", "304", "240", "~
## $ company                         <chr> "NULL", "NULL", "NULL", "NULL", "NULL",~
## $ days_in_waiting_list            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ customer_type                   <chr> "Transient", "Transient", "Transient", ~
## $ adr                             <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00,~
## $ required_car_parking_spaces     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ total_of_special_requests       <dbl> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, ~
## $ reservation_status              <chr> "Check-Out", "Check-Out", "Check-Out", ~
## $ reservation_status_date         <date> 2015-07-01, 2015-07-01, 2015-07-02, 20~
```

Question 1 – Quantitative Outcome Variable

In this part of the assignment, your outcome variable is `adr`, the Average Daily Rates in Euros, which represents the average revenue per occupied room night across the length of the hotel stay. You may choose any of the following variables as your primary covariate: `lead_time`, `hotel`, or `stays_in_week_nights`. [See original assignment document for each subpart, in detail.]

Solution

(a) Let `lead_time` be the primary covariate of interest, defined as the number of days between when the booking reservation was made and when the booking was honored (arrival day). Let `hotel` be the interaction term, which is a categorical variable describing one of the two hotels for which data was collected. Lead time is plausibly related to pricing (which `adr` represents in this case) via dynamic pricing: customers who book further in advance may see different pricing strategies than those who book much closer to when they wish to be at the hotel. This relationship between `lead_time` and `adr` may also depend on the type of hotel (resort versus city), since pricing strategies and demand patterns are likely to be influenced by these differing types. For example, city hotels may exhibit more stable prices, driven by a more steady demand of business or corporate travel whereas resort hotels may see stronger seasonal booking effects, driven by tourist demands.

Let β_1 correspond to the effect of `lead_time`, which we encode as x_1, β_2 correspond to the effect of the hotel type (where we set the default to be "resort"), which we encode as x_2, and let β_3 correspond to the effect of the interaction between `lead_time` and `hotel`. For the main effect, we test

$$H_0^m : \beta_1 = \beta_3 = 0 \qquad \text{vs.} \qquad H_A^m : \text{at least one of } \beta_1, \beta_3 \neq 0$$

which is to say that our null hypothesis is there is no association between lead time and ADR, regardless of the hotel. Under the assumption this null is true, the coefficients for the two hotels would both be equal to zero: the first coefficient would be β_1 whereas the second would be $\beta_1 + \beta_3$, and both terms are assumed to be equal to 0. For our interaction, we test

$$H_0^i : \beta_3 = 0 \qquad \text{vs.} \qquad H_A^i : \beta_3 \neq 0,$$

which is to say that our null is that any relationship between `lead_time` and `adr` does not differ by hotel type and that our alternative is that any effect of `lead_time` on `adr` does depend on the hotel type. Thus, our full model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon,$$

where β_0 corresponds to the baseline `adr` for the city hotel when `lead_time = 0`, (recall $x_2 = \mathbb{1}\{\text{hotel type} = \text{"resort"}\}$), β_1 corresponds to the effect of `lead_time` for the city hotel, β_2 corresponds to the difference in baseline `adr` levels between hotels, β_3 corresponds to the difference in `lead_time` effect between hotels, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ corresponds to noise.

[Note: for the rest of the assignment, we assume the noise in our model is negligible and that it can be ignored.]

(b) Before looking at any of the data, we anticipate that ADR, which corresponds to prices, will likely be right-skewed. This is because most of the hotels will likely exhibit average daily rates that are close to some "reasonable" price (otherwise the hotel would not have many reservations and would not have any business), with a few exceptions that yield a very high ADR. To address this, consider we redefine our outcome variable to be

$$y = \log(\text{adr} + 1).$$

Applying a logarithmic transformation means we can reduce the skewness, stabilize the variance, and make it easier to understand the nature of the relationships between our variables. We add 1 to the log values to handle any zeroes ADR values. Our primary model keeps its same form (ignoring the error term):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 = \beta_0 + \beta_1 x_1 + \beta_2 \cdot \mathbb{1}\{\text{hotel} = \text{`Resort'}\} + \beta_3 x_1 \cdot \mathbb{1}\{\text{hotel} = \text{`Resort'}\}.$$

[Note: if you decide a transformation is not going to be included in your primary model, you will get credit for this part as long as you justify your choice.]

Below is some R code to provide some summary statistics for our variables:

```
hotel_bookings %>%
  summarize(
    adr_mean = mean(adr, na.rm = TRUE),
    adr_sd   = sd(adr, na.rm = TRUE),
    adr_min  = min(adr, na.rm = TRUE),
    adr_max  = max(adr, na.rm = TRUE),

    log_adr_mean = mean(log(adr + 1), na.rm = TRUE),
    log_adr_sd = sd(log(adr + 1), na.rm = TRUE),
    log_adr_min = min(log(adr + 1), na.rm = TRUE),
    log_adr_max = max(log(adr + 1), na.rm = TRUE),
  )
```

```
## # A tibble: 1 x 8
##   adr_mean adr_sd adr_min adr_max log_adr_mean log_adr_sd log_adr_min
##      <dbl>  <dbl>   <dbl>   <dbl>        <dbl>      <dbl>       <dbl>
## 1     102.   50.5   -6.38    5400         4.48      0.735           0
## # i 1 more variable: log_adr_max <dbl>
```

```
cat("Number of NaNs afer transformation:", sum(is.nan(log(hotel_bookings$adr + 1))))
```

```
## Number of NaNs afer transformation: 1
```

```
hotel_bookings %>%
  summarize(
    lt_mean = mean(lead_time, na.rm = TRUE),
    lt_sd   = sd(lead_time, na.rm = TRUE),
    lt_min  = min(lead_time, na.rm = TRUE),
    lt_max  = max(lead_time, na.rm = TRUE)
  )
```

```
## # A tibble: 1 x 4
##   lt_mean lt_sd lt_min lt_max
##     <dbl> <dbl>  <dbl>  <dbl>
## 1    104.  107.      0    737
```

To produce these summary statistics, we explicitly chose to not consider any rows for which there are `NA` values. While this could potentially mean we lose some information if a row only has only one `NA` value, this procedure allows for consistency across all summary statistics we wish to report. Applying the logarithmic transformation on `adr` actually only yields one invalid term, namely the one corresponding to `adr_min`, which was negative to begin with. We make a design choice to ignore that one row in our analysis. Indeed, with a sample of size 119,390, this one row that produces a `NA` value after the transformation is unlikely to markedly change our analysis/results.

[Note: alternative option would have been to apply a transformation that shifts everything up by this minimum value. However, since we applied the transformation **WITHOUT** looking at our data, we leave it as is.]

Some observations include that the logarthmic distribution for ADR seems to be much more reasonable than the raw values and that `lead_time` also appears to exhibit some skewness (evidenced by the fact that `lt_max` is 737 days). This likely corresponds to many short-notice bookings and a more spread-out tail of early reservations.

Since `hotel` is a binary variable, its summary statistics can be captured by the proportion of rows for which each hotel type appears:

```
table(hotel_bookings$hotel)
```

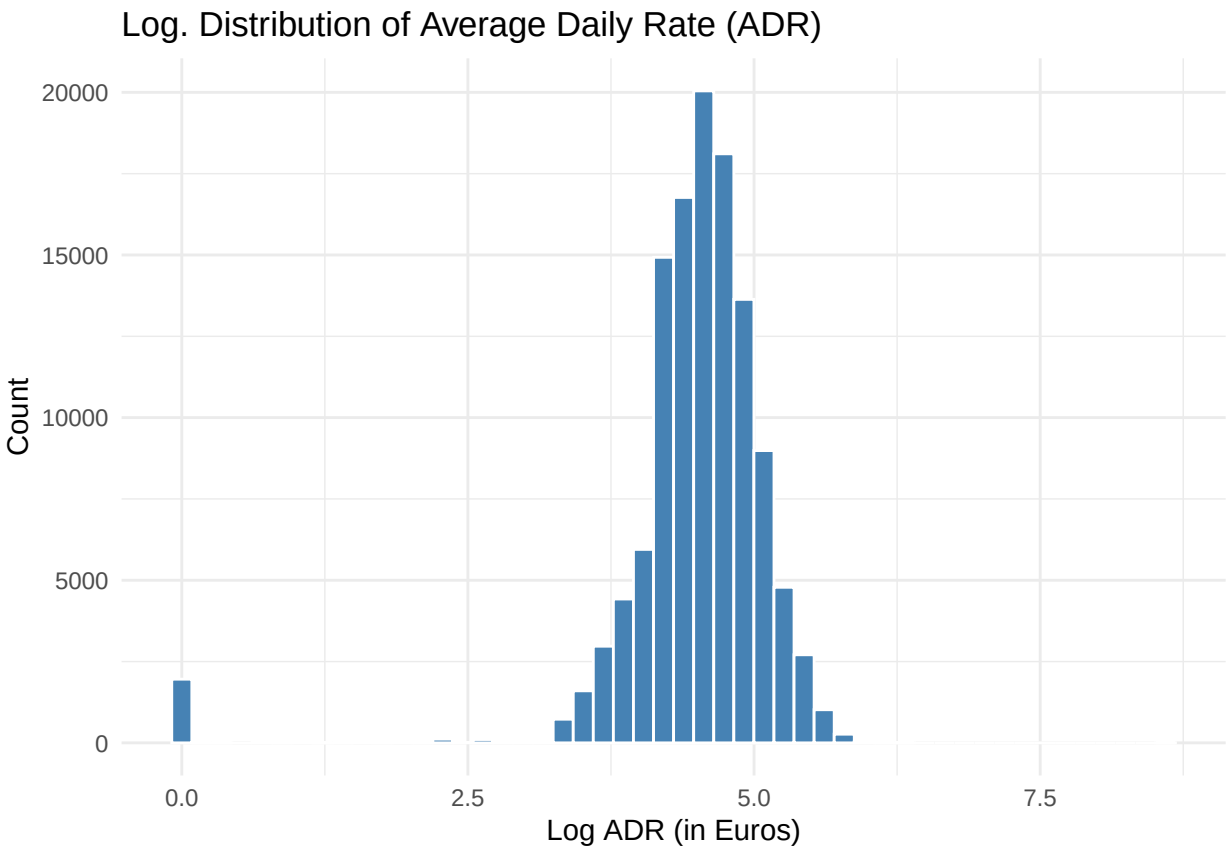
```
##
##   City Hotel Resort Hotel
##        79330        40060
```

```
prop.table(table(hotel_bookings$hotel))
```

```
##
##   City Hotel Resort Hotel
##     0.664461     0.335539
```

The hotel type classes are not exorbitantly imbalanced: a ratio of roughly 2:1 seems reasonable, especially considering that a city likely attracts a wider range of prospective visitors than a resort hotel.We can analyze the distribution of ADR, after applying the transformation, more closely:

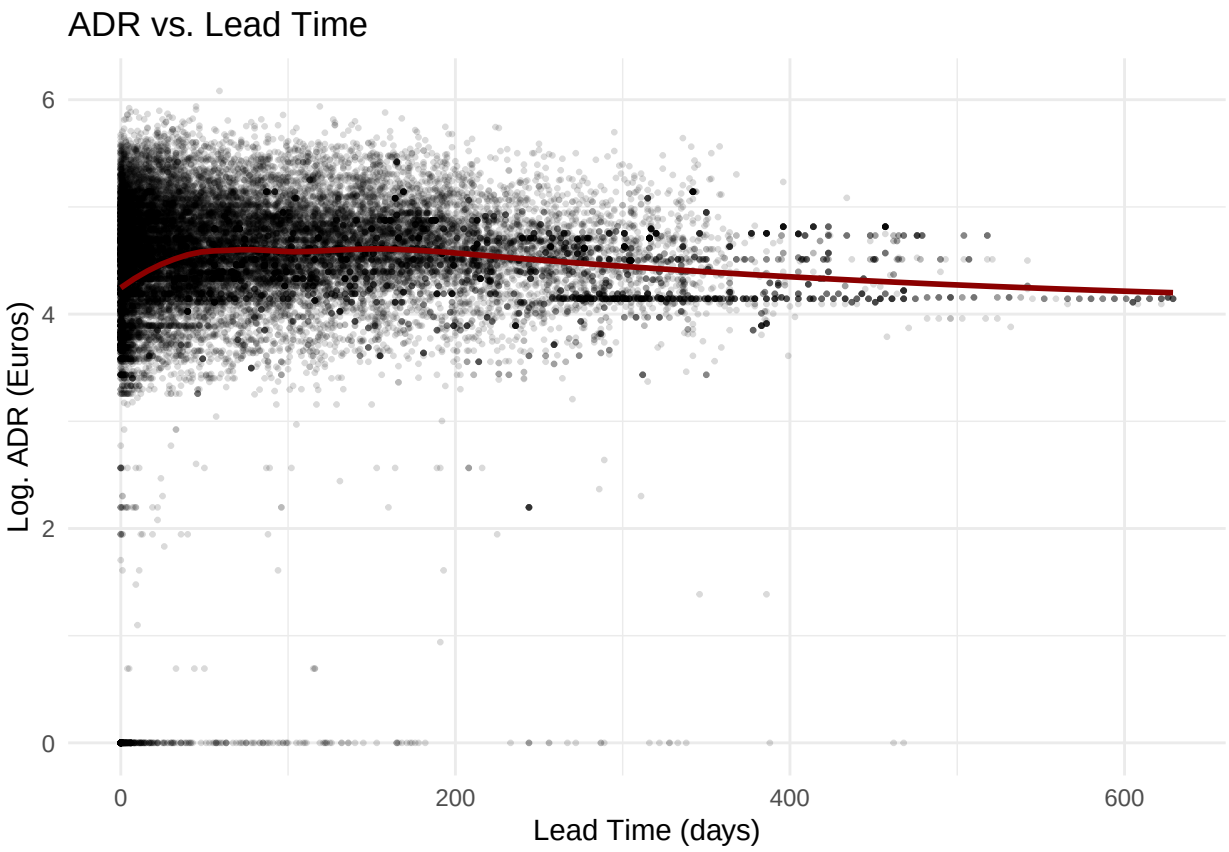
```
ggplot(hotel_bookings, aes(x=log(adr + 1))) +
  geom_histogram(bins=50, fill='steelblue', color='white') +
  labs(title="Log. Distribution of Average Daily Rate (ADR)",
       x="Log ADR (in Euros)", y="Count") +
  theme_minimal()
```



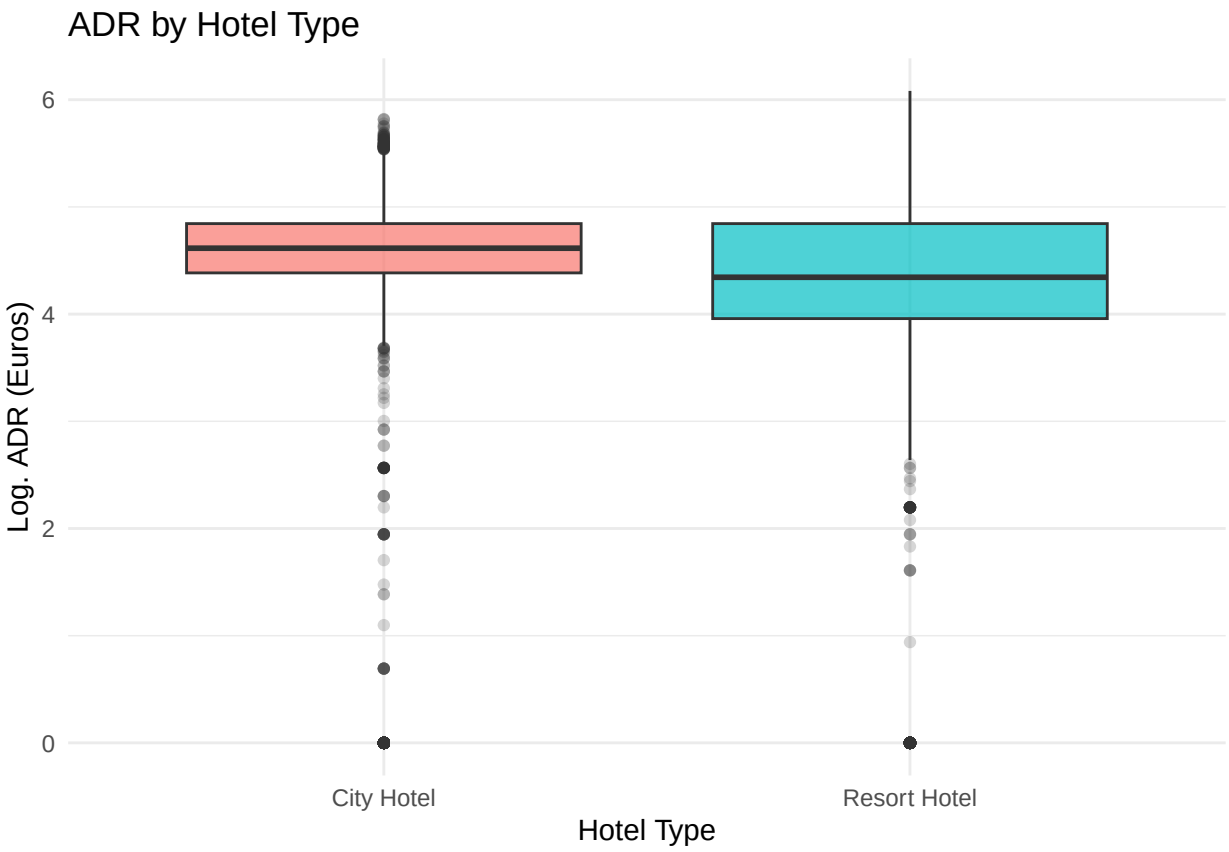
Our choice of transformation for the ADR values seems to be helpful, since we get a distribution that appears to be much more normal in shape. Sure, there are some values hovering around 0, but this plot seems to suggest that we could benefit from looking at our model using this transformation.

Now we analyze the relationships between the various variables in our model. To speed up computations and our plotting, we select a subset of our data. That is, instead of visualizing nearly 120,000 data points, we select about 25% for our visualization. (It is unlikely that the distribution will change too much with this subset.)

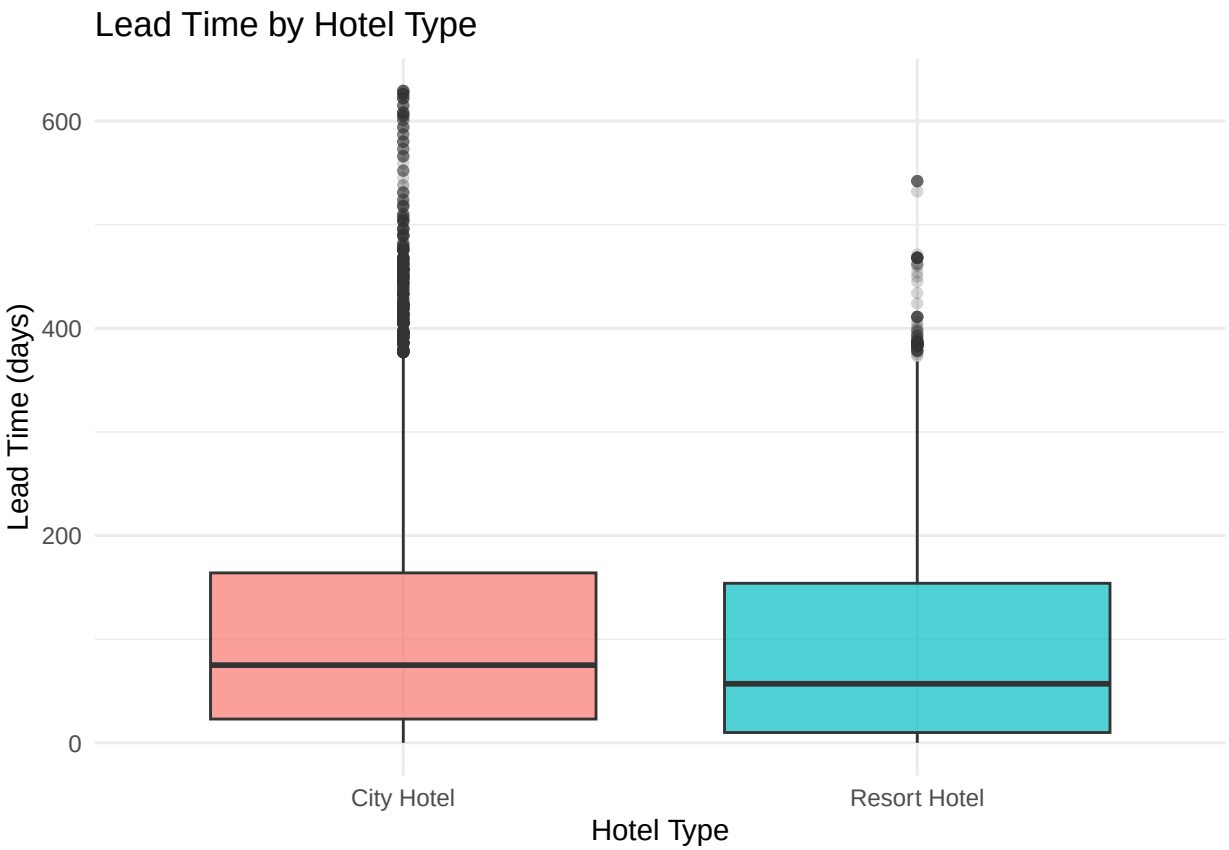
```
# Set seed for reproducibility in plotting data
plotting_data <- hotel_bookings %>% sample_frac(0.25) # Get 25% of data for plotting

ggplot(plotting_data, aes(x = lead_time, y=log(adr + 1))) +
  geom_point(alpha=0.15, size=0.5) +
  geom_smooth(method='loess', color='darkred', se=FALSE) +
  labs(title="ADR vs. Lead Time", x="Lead Time (days)", y="Log. ADR (Euros)") +
  theme_minimal()
```




```
ggplot(plotting_data, aes(x=hotel, y=log(adr + 1), fill=hotel)) +
  geom_boxplot(alpha=0.7, outlier.alpha=0.2) +
  labs(title="ADR by Hotel Type", x="Hotel Type", y="Log. ADR (Euros)") +
  theme_minimal() +
  theme(legend.position='none')
```




```
ggplot(plotting_data, aes(x=hotel, y=lead_time, fill=hotel)) +
  geom_boxplot(alpha=0.7, outlier.alpha=0.2) +
  labs(title="Lead Time by Hotel Type", x="Hotel Type", y="Lead Time (days)") +
  theme_minimal() +
  theme(legend.position="none")
```

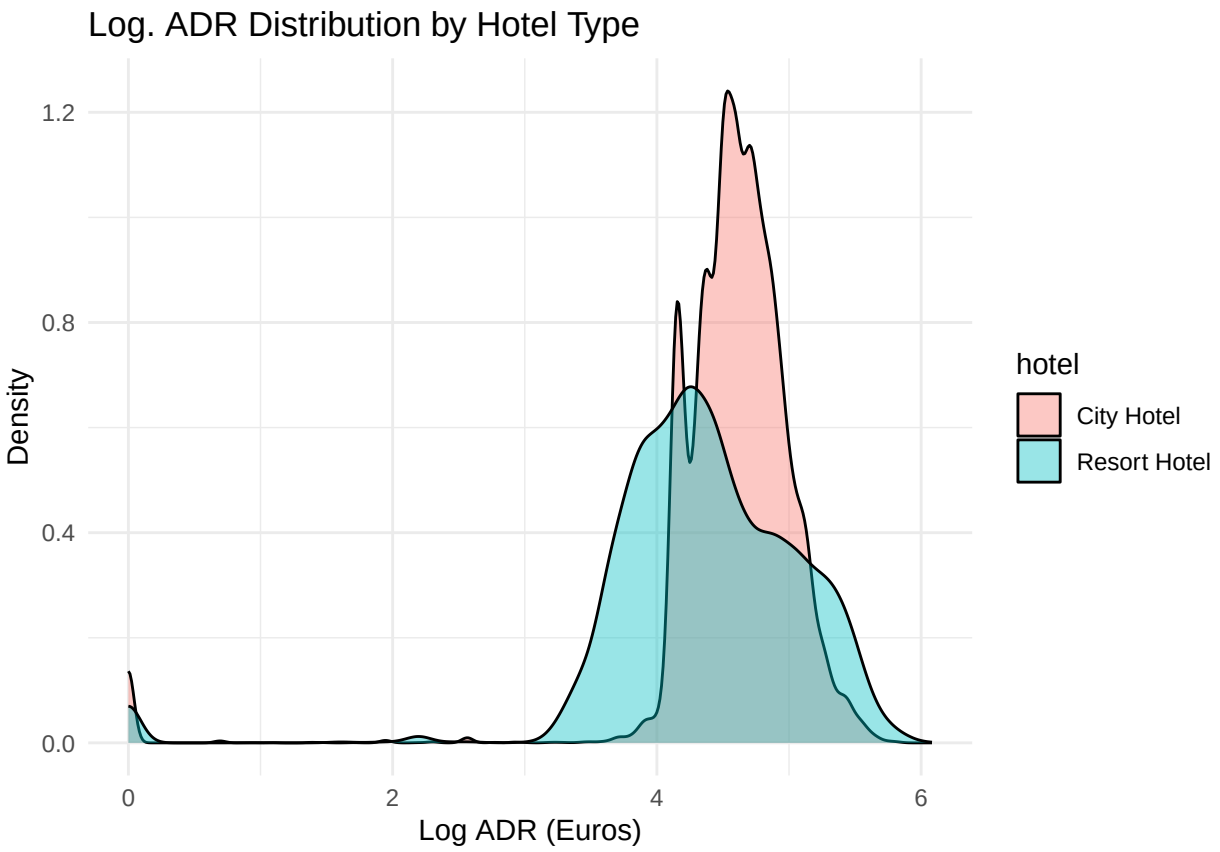


These plots yield a couple of interesting observations:

1. There appears to be no discernible, entirely linear trend between our transformed ADR values and lead time. Relatedly, we see a number of data points whose transformed ADR value is 0, regardless of the number of lead days.
2. After applying our transformation to the ADR values, the city hotel seems to have a slightly higher median logarithmic ADR value and its spread appears to be reduced (as viewed by its interquartile range). Having said that, the city hotel still seems to have a wider range, with more extreme values than the resort hotel.
3. Lead time does not seem to be influenced by hotel type very much. Both hotels produce similar boxplots, with about the same median and interquartile range.

We check out one more plot in case this density plot reveals any more information about relationship between hotel type and ADR:

```
ggplot(plotting_data, aes(x=log(adr + 1), fill=hotel)) +
  geom_density(alpha = 0.4) +
  labs(title="Log. ADR Distribution by Hotel Type", x="Log ADR (Euros)", y="Density") +
  theme_minimal()
```



This distributional plot indicates that the resort hotel exhibits ADR values that are more spread out (hence why the its peak density is a lower than that of the city hotel). All in all, we might say that there are some noticeable differences across hotel types when it comes to our outcome variable, and our transformation seemed to have helped with that.

(c) Recall our two sets of hypotheses, for the main effect and the interaction, respectively, were

$$H_0^m : \beta_1 = \beta_3 = 0 \text{ vs. } H_A^m : \text{at least one of } \beta_1, \beta_3 \neq 0 \qquad \text{and} \qquad H_0^i : \beta_3 = 0 \text{ vs. } H_A^i : \beta_3 \neq 0.$$

For our main effect, we employ a partial $\mathcal{F}$-test. The model defined above, which we will refer to in the code below as `q1_model`, acts as our full model, and our "reduced" (or null) model assumes `lead_time` does not feature. Thus, our null model is of the form `log(adr + 1) ~ hotel`; we refer to this model moving forward as `q1_reduced_model`. We create this model and compare using an analysis of variances:

```
q1_reduced_model <- lm(log(adr + 1) ~ hotel, data=hotel_bookings)
q1_model <- lm(log(adr + 1) ~ lead_time * hotel, data=hotel_bookings)
anova(q1_reduced_model, q1_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(adr + 1) ~ hotel
## Model 2: log(adr + 1) ~ lead_time * hotel
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1 119387 63137
## 2 119385 62930  2    207.23 196.57 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the produced output, our F-statistic is 196.57 and its corresponding p-value is approximately

zero, which means we should reject H_0^m at the significance level $\alpha = 0.05$. There is statistically significant evidence that lead time does contribute to predicting log-transformed ADR in at least one hotel category; said differently and more in terms of the models for our $\mathcal{F}$-test, including `lead_time` and its interaction with the type of hotel significantly improves the explanatory power of our linear model relative to a model that looks at hotel type alone. To interpret this effect size, we can understand how much more of the variation in `log(adr + 1)` in this full model can be attributed to adding `lead_time` and its interaction with `hotel`, as compared to `hotel` on its own. This would involve computing the R^2 value from the ANOVA table above:

$$R^2_{\text{partial}} = \frac{\text{RSS}_r - \text{RSS}_f}{\text{RSS}_r} = \frac{207.23}{63137} \approx 0.0033.$$

Here, RSS_r and RSS_f denote the residual sum of squares for our "reduced" and full models, respectively. We have obtained a statistically significant but practically modest result: adding lead time and its interaction with hotel explains approximately 0.33% of the residual variation in log-transformed ADR beyond what the hotel type explains alone.

For the interaction term and H_0^i, it suffices to just observe the output of `q1_model`.

```
summary(q1_model)
```

```
##
## Call:
## lm(formula = log(adr + 1) ~ lead_time * hotel, data = hotel_bookings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7854 -0.2283  0.0466  0.3283  4.0327
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                   4.566e+00  3.626e-03 1259.345  < 2e-16 ***
## lead_time                    -1.200e-04  2.323e-05   -5.165 2.41e-07 ***
## hotelResort Hotel            -3.066e-01  6.184e-03  -49.570  < 2e-16 ***
## lead_time:hotelResort Hotel  8.338e-04  4.393e-05   18.979  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.726 on 119385 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.02452,    Adjusted R-squared:  0.02449
## F-statistic:  1000 on 3 and 119385 DF,  p-value: < 2.2e-16
```

From the above summary, we get $\hat{\beta}_3 = 0.000834$ and its corresponding p-value is $p \approx 0$, which means we would reject H_0^i at the significance level $\alpha = 0.05$. There is statistically significant evidence that the association between lead time and ADR (under our chosen logarithmic transformation) differs by hotel type. Practically, this means that for resort hotels, longer lead time is associated with a $\hat{\beta}_1 + \hat{\beta}_3 \approx 0.00714$ increase in `log(adr + 1)`, or that earlier bookings are associated with higher-priced stays at the resort hotel. To interpret this effect size via percent change, we have the relation

$$\%\Delta\text{ADR} \approx (e^{\beta} - 1) \cdot 100,$$

where β denotes the coefficient of the effect term we are interested. We compute this percent change for each:

```r
lt_effect_city_pct <- (exp(summary(q1_model)$coefficients[2,1]) - 1) * 100
lt_effect_city_pct
```

```
## [1] -0.01199919
```

```r
lt_effect_resort_pct <- (
  exp(summary(q1_model)$coefficients[4,1] + summary(q1_model)$coefficients[2,1])
  - 1 ) * 100
lt_effect_resort_pct
```

```
## [1] 0.07140505
```

```r
interaction_effect_pct <- (exp(summary(q1_model)$coefficients[4,1]) - 1) * 100
interaction_effect_pct
```

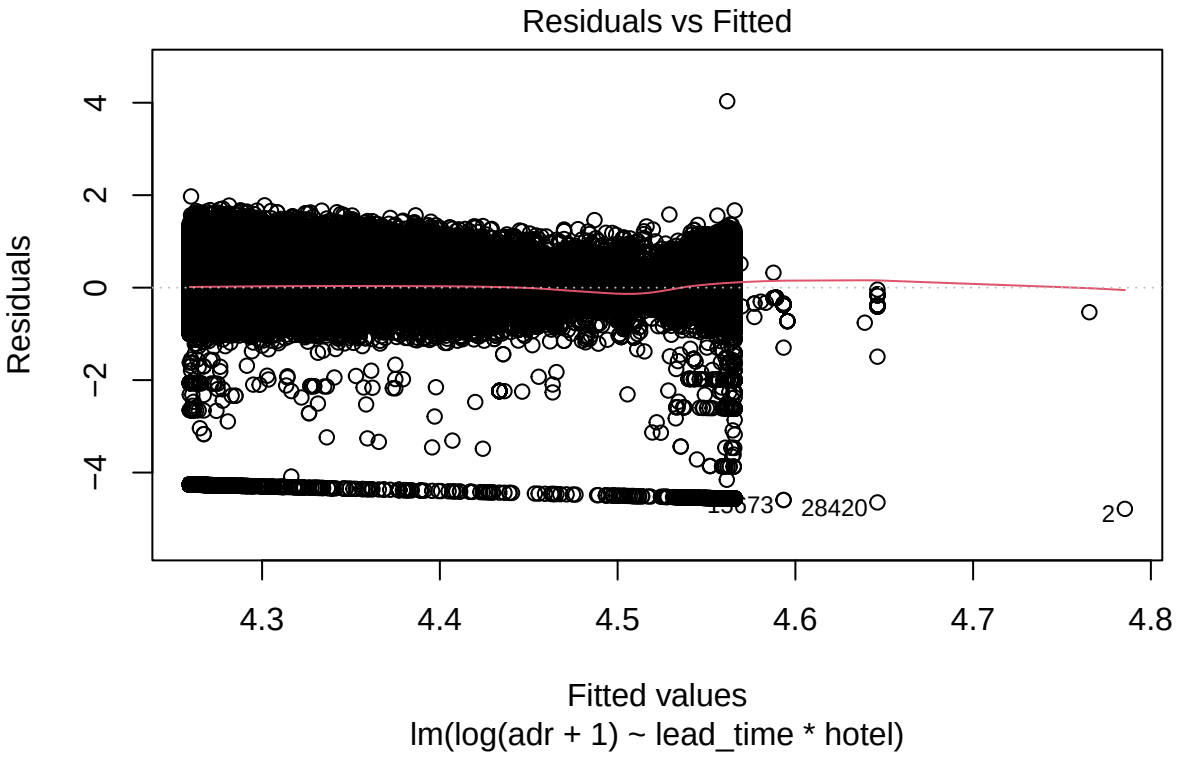
```
## [1] 0.08341426
```

That is, a one-day increase in lead time is associated with a roughly -0.012% change in ADR for city hotels while a one-day increase in lead time is associated with approximately a 0.071% change in ADR for resort hotels. Both are relatively small percent changes, but they are statistically significant given the large sample size. Similarly, the effect of lead time on ADR differs between the two hotels by approximately 0.083% per day, meaning resort hotels show a more positive relationship between booking horizon and price as compared to city hotels.

The overall conclusion is that although the magnitude of per-day effects appears to be relatively small in absolute terms, they are highly statistically significant the oppositely directional slopes highlight a lack of homogeneity with regards to pricing strategies across the hotels.

(d) We start by checking the four conditions for multiple linear regression, as described in Lecture 10. For linearity, we check a plot of the residuals versus the fitted values since this helps assess whether the relationship between the predictors and the outcome is approximately linear.

```r
plot(q1_model, which=1)
```



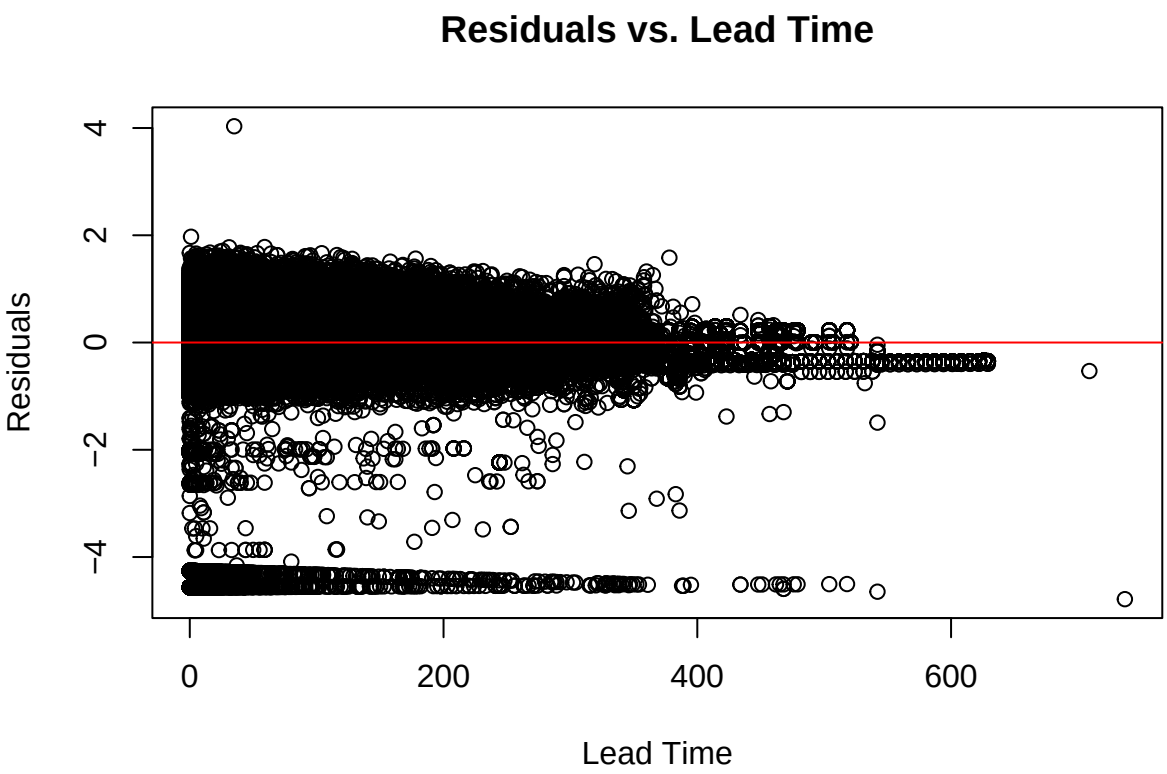

```
mean(q1_model$residuals)
```

```
## [1] -8.436857e-16
```

After applying the log transformation to ADR, the residuals-versus-fitted plot does not indicate any major violation of linearity. The residuals appear randomly scattered around zero across the range of fitted values, with no strong systematic curvature evident. Additionally, the residuals remain centered near zero throughout the plot, supporting the appropriateness of the linear functional form. While there is a noticeable concentration of residuals near -4, this is likely a consequence of the log transformation compressing observations with ADR values near zero, rather than evidence of nonlinearity in the model.

To check homoskedasticity (equal variances), we plot the residuals against our primary covariate, `lead_time` since this helps assess whether the variance of our residuals is constant across the key predictor values.

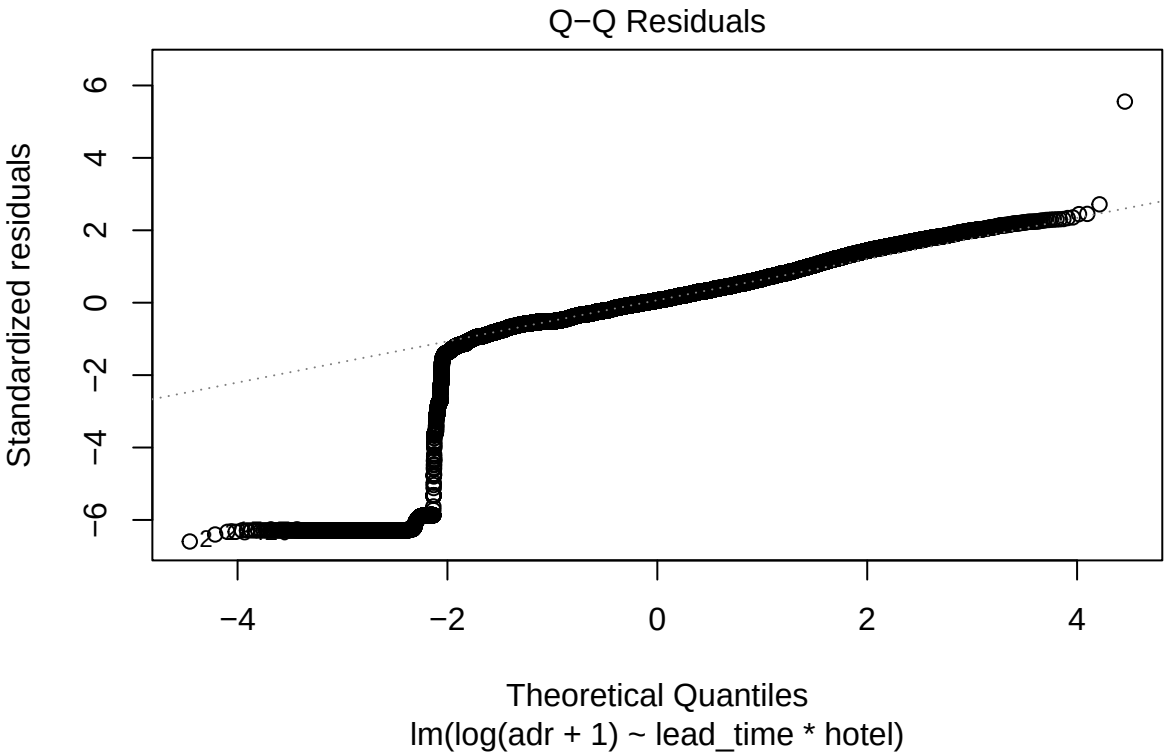
```
plot(model.frame(q1_model)$lead_time, resid(q1_model), xlab="Lead Time", ylab="Residuals",
     main="Residuals vs. Lead Time")
abline(h=0, col='red')
```



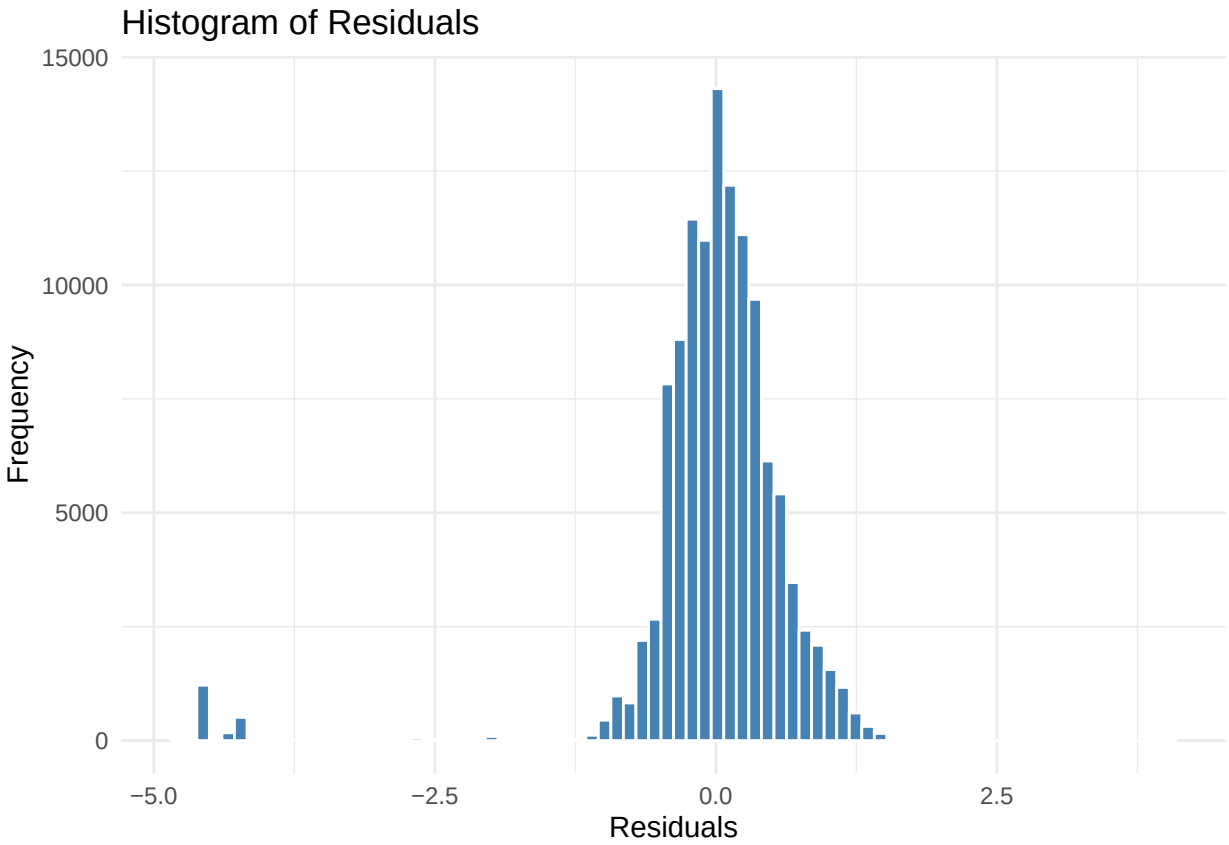
This residual plot suggests some deviation from homoskedasticity, despite our chosen logarithmic transformation applied to ADR. In particular, the spread of the residuals appears to be larger for smaller lead times and narrows as lead time increases, an indication that residual variance is not perfectly constant across the entire predictor range. However, we should note that our chosen transformation substantially improves variance stability relative to the raw ADR scale, and the observed deviations would not be considered severe enough to undermine any inference made using this linear model. The main reason for this is the size of our sample: with nearly 120,000 observations, some deviations are bound to exist, though they may not directly change our faith in the model's validity. It is also worth mentioning the horizontal band of residuals near -4, which is largely attributable to observations whose original ADR value was near zero. We would not consider these as indicating a violation of homoskedasticity but rather consequences of our chosen transformation.

To check normality, we will check a Q-Q plot and the histogram of the residuals, since both show will indicate some kind of deviation in the tails of the distribution of the residuals.

```
plot(q1_model, which=2)
```



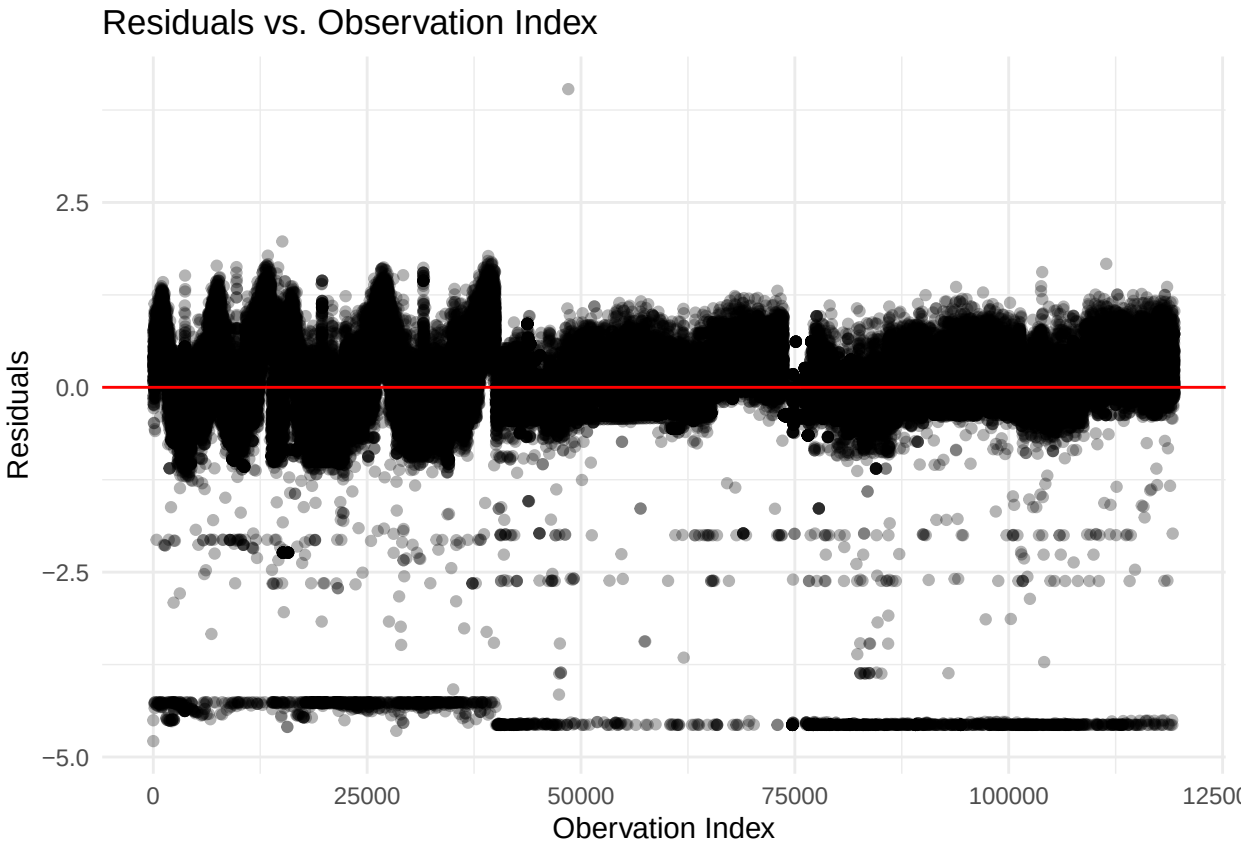

```
ggplot(data.frame(resid = resid(q1_model)), aes(x = resid)) +
  geom_histogram(bins = 80, fill = "steelblue", color = "white") +
  labs(title="Histogram of Residuals", x="Residuals", y="Frequency") +
  theme_minimal()
```



The Shapiro-Wilks test does not apply directly on this dataset given its size. [Note: it is possible to take a random sample of size $n = 5000$ and run this test on that sample. You will not lose points for not doing the test in this setting, as long as there is some mention of the test.] The Q-Q plot and the residual histogram suggest mild departures from normality over the entire dataset, particularly at the lower tail of the distribution. This deviation is likely driven in part by our transformation, which compressed the ADR values that were already near zero and yielded a cluster of low residual values. Nevertheless, the residual distribution is largely symmetric overall, and, given the large sample size, inference on the regression coefficients remains reliable due to the asymptotic normality of ordinary least-squares estimators implied by the Central Limit Theorem; this can hold despite the residuals themselves not being perfectly normally distributed.

Lastly, we check for any potential violation of independence by plotting the residuals against time, ordering our observations by index.

```
df_q1_model <- model.frame(q1_model)
ggplot(data.frame(index=1:nrow(df_q1_model), resids=resid(q1_model)),
       aes(x=index, y=resids)) +
  geom_point(alpha=0.3) +
  geom_hline(yintercept=0, col='red') +
  labs(title="Residuals vs. Observation Index", x="Obervation Index", y="Residuals") +
  theme_minimal()
```



The residuals-versus index plot here does not suggest strong evidence against the independence assumption. For the most part, the residuals remain roughly centered about zero through the observation ordering, with no "blatantly obvious" long-range trends or sequential structure. There are some localized changes in the spread in the main cluster of residuals (e.g. the upward-trending swells around index 2500 or the longer trend between indices 50000 and 75000): these might have some seasonal trend component that is not immediately apparent from looking at the dataset as a whole and would war-

rant further exploration. There are also some interesting cluster-based residuals, most noticeable in the four seemingly-stand-out horizontal "beams" in the data. These "beams" (the main one about 0 and the lowest one, around -4, are easy to see) are likely a reflection of heterogeneity in hotel categories or booking characteristiccs rather than any clear autocorrelation. Moreover, because the dataset consists primarily of distinct booking events rather than repeated longitudinal measurements on the same unit, the assumption of independence is reasonably plausible in this setting.

Having found that none of the assumptions for multiple regression are *grossly or seriously* violated, we move on to fitting an alternative model. Consider we fit a reduced model, where there is no interaction term. That is, we want to fit a model of the form `log(adr + 1) ~ lead_time + hotel`. In addition to not having the interaction term, this alternative model is more restrictive in that it does not allow for heterogeneous slopes between `log(adr + 1)` and `lead_time` based on the hotel type; here, we have reduced flexibility by constraining the slopes to be constant across both groups. Moreover, the interpretation is that of average effect of `lead_time` on `log(adr + 1)` across hotels, not hotel-specific `lead_time` effects. As a result of this additive structure, our model becomes:

$$\tilde{y} = \log(\text{adr} + 1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \beta_0 + \beta_1 \cdot \text{lead\_time} + \beta_2 \cdot \mathbb{1}\{\text{hotel} = \text{“Resort”}\}.$$

Since there is no interaction term present in this model, we ignore the test for it and instead just test $H_0^a : \beta_1 = 0$ versus $H_A^a : \beta_1 \neq 0$. The R code below fits this reduced model and produces its output:

```
q1_alt_model <- lm(log(adr + 1) ~ lead_time + hotel, data=hotel_bookings)
summary(q1_alt_model)
```

```
##
## Call:
## lm(formula = log(adr + 1) ~ lead_time + hotel, data = hotel_bookings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5989 -0.2184  0.0495  0.3279  4.0501
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         4.540e+00  3.371e-03 1347.016  < 2e-16 ***
## lead_time           1.132e-04  1.975e-05    5.731 9.99e-09 ***
## hotelResort Hotel -2.253e-01  4.469e-03  -50.409  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7271 on 119386 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.02157,    Adjusted R-squared:  0.02156
## F-statistic:  1316 on 2 and 119386 DF,  p-value: < 2.2e-16
```

Based on the results for `lead_time`, we get $\hat{\beta}_1 = 0.000113$ and its corresponding p-value is $p \approx 9.99 \cdot 10^{-9}$. Thus, we would reject the null hypothesis at the $\alpha = 0.05$ significance level. There is statistically significant evidence to suggest that `lead_time` has an association on log-transformed ADR after adjusting for hotel type. Practically, this means that holding the hotel category constant, each additional day of lead time is associated with a $(e^{\widehat{\beta}_1} - 1) \cdot 100 \approx 0.011\%$ increase in ADR each day. (We cannot make a claim about the interaction effect here since no such term is present in our model.) Interestingly enough, the hotel coefficient remains negative, which means that the resort hotel (as indicated by the output) tends to have a lower `log(adr + 1)` value relative to city hotels, when the lead time is fixed.

The primary conclusion that `lead_time` does have an association with log-transformed ADR values remains consistent across both of our models. However, a key difference emerges in the interpretation of `lead_time`: in the primary model, the effect of lead time different substantially for hotel type, being slightly negative for city hotels and positive for resort hotels. In this alternative (interaction-less) model, the estimated lead time effect becomes positive, and is actually less informative than the values we observed in our primary analysis. This is because the alternative model forces a single, common slope across both hotel categories, meaning that this effect of `lead_time` is pooled across both groups and masks the heterogeneity we found in the results of our primary analysis. Also, the alternative model exhibits slightly a lower R^2 value (0.0215743 versus 0.0245173) as compared to our full model, which suggests that omitting the interaction term yields slightly lower explanatory power in the model. The difference is not very big, but it is something perhaps worth noting. Overall, though, the main conclusion stated at the opening of this paragraph holds.

Despite the general robustness in the findings, there are a couple of limitations to these models. First, their simplicity (no more than two predictors were considered in each) lead to low explanatory power ($R^2 \approx 0.02$), which is to suggest that most of the variation in ADR is driven by other factors not included in either of these models. Second, although there were valid reasons for which the log-transformed ADR values were more appropriate for our model than the raw ADR values, even with the transformation, we notice small deviations from normality are observed. Lastly, since our dataset is entirely observational (we assume that the data was simply recorded as time passed and that there was no well-established procedure with which the data would then be collected), it would not be reasonable to say that the relationships explored in either our primary or alternative models indicate a causal relationship; instead, we must make due with simply saying there exist some associations between the variables.

Question 2 – Binary Output Variable

In this part of the assignment, your outcome variable is `is_canceled`, a binary indicator variable that identifies whether the booking was ultimately canceled (encoded as a 1) or honored (encoded as a 0). You may choose any of the following as your primary covariate of interest: `lead_time`, `previous_cancellations`, or `deposit_type`. [See original assignment document for each subpart, in detail.]

Solution

(a) Let `lead_time` be the primary covariate of interest, and let `deposit_type` be our confounder. This predictor is a reasonable choice, with respect to our outcome variable `is_canceled`, because there could be some nuanced relationship between when bookings were made and if they ended up being cancelled; for example, one possible relationship might suggest bookings made farther in advance are more likely to be cancelled since there is more time for customers to change or revise their travel plans. The choice of `deposit_type` as our confounder is also a reasonable one, since customers who pay a non-refundable deposit may be less likely to cancel their reservation, and there may be requirements associated with the kind of deposit made based on how soon/late the booking was made. Not taking into account the possible effect of `deposit_type` could have introduced potential biases in the estimated relationship between `lead_time` and the cancellation probability.

For our model, let

$$Y_i = \begin{cases} 1, & \text{if booking } i \text{ was cancelled,} \\ 0, & \text{otherwise} \end{cases},$$

and let x_i be the observed lead time of booking i. (We denote the random variable X_i to refer to the lead time of booking i, and then x_i corresponds to a specific, observed instance of that random variable.) Since `deposit_type` has 3 distinct categories, we will use two indicator variables. If "No Deposit" is the reference category, let

$$D_{1i} = \mathbb{1}\{\texttt{deposit\_type} = \text{“Refundable”}\} = \begin{cases} 1, & \text{if } \texttt{deposit\_type} \text{ is “Refundable”,} \\ 0, & \text{otherwise} \end{cases}$$

and

$$D_{2i} = \mathbb{1}\{\texttt{deposit\_type} = \text{“Non-Refund”}\} = \begin{cases} 1, & \text{if } \texttt{deposit\_type} \text{ is “Non-Refund”,} \\ 0, & \text{otherwise} \end{cases}$$

Then, our primary logistic regression model is

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot X_i + \beta_2 D_{1i} + \beta_3 D_{2i},$$

where $p_i = \mathbb{P}(Y_i = 1 \mid X_i = x_i)$ denotes the probability that booking i was cancelled given a lead time of x_i days. Since β_1 is our primary parameter of interest, we test

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_A : \beta_1 \neq 0.$$

That is, our null hypothesis says there is no association between the lead time of a booking and the odds that it gets cancelled, after controlling for deposit type, while our alternative says there is some association between lead time and cancellation odds, after controlling for deposit type.

(b) Based solely on the variable name (`is_canceled`) and on the context in reference to hotel bookings, we would expect most of the hotels bookings to be honored. While there are bound to be cancellations,

most hotel reservations are eventually carried through, meaning that we would expect the dataset to contain more 0s than 1s. We get the true class distribution as follows:

```
table(hotel_bookings$is_canceled)              # Get class counts
```

```
##
##     0     1
## 75166 44224
```

```
prop.table(table(hotel_bookings$is_canceled))  # Get class proportions
```

```
##
##         0         1
## 0.6295837 0.3704163
```

These produced results suggest 63% of the hotel bookings were honored, which seems to be in line with our intuition that there would be more honored bookings than cancelled ones. Although there is a class imbalance (about 13% off-center), this imbalance is not severe enough to cause issues fitting a logistic regression model. Moreover, the number of observations in each class is sizable, meaning our model should still be able to correctly estimate the coefficients without the need for any other alterations to our model.

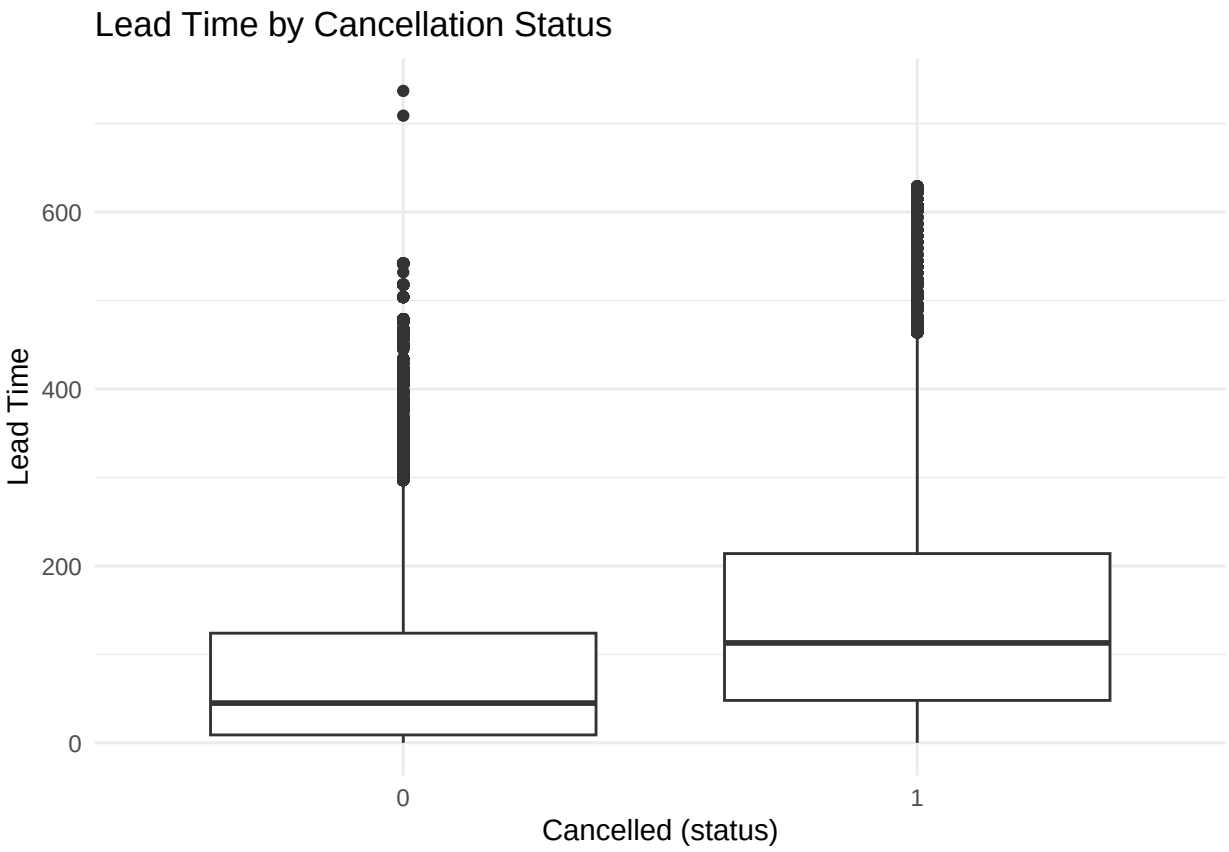
Since `lead_time` is a quantitative predictor, we can easily stratify our summary statistics by `is_canceled`, which we do as follows:

```
hotel_bookings %>%
  group_by(is_canceled) %>%
  summarize(
    n = n(),
    lt_mean = mean(lead_time, na.rm = TRUE),
    lt_median = median(lead_time, na.rm = TRUE),
    lt_sd = sd(lead_time, na.rm = TRUE),
    lt_min = min(lead_time, na.rm = TRUE),
    lt_max = max(lead_time, na.rm = TRUE)
  )
```

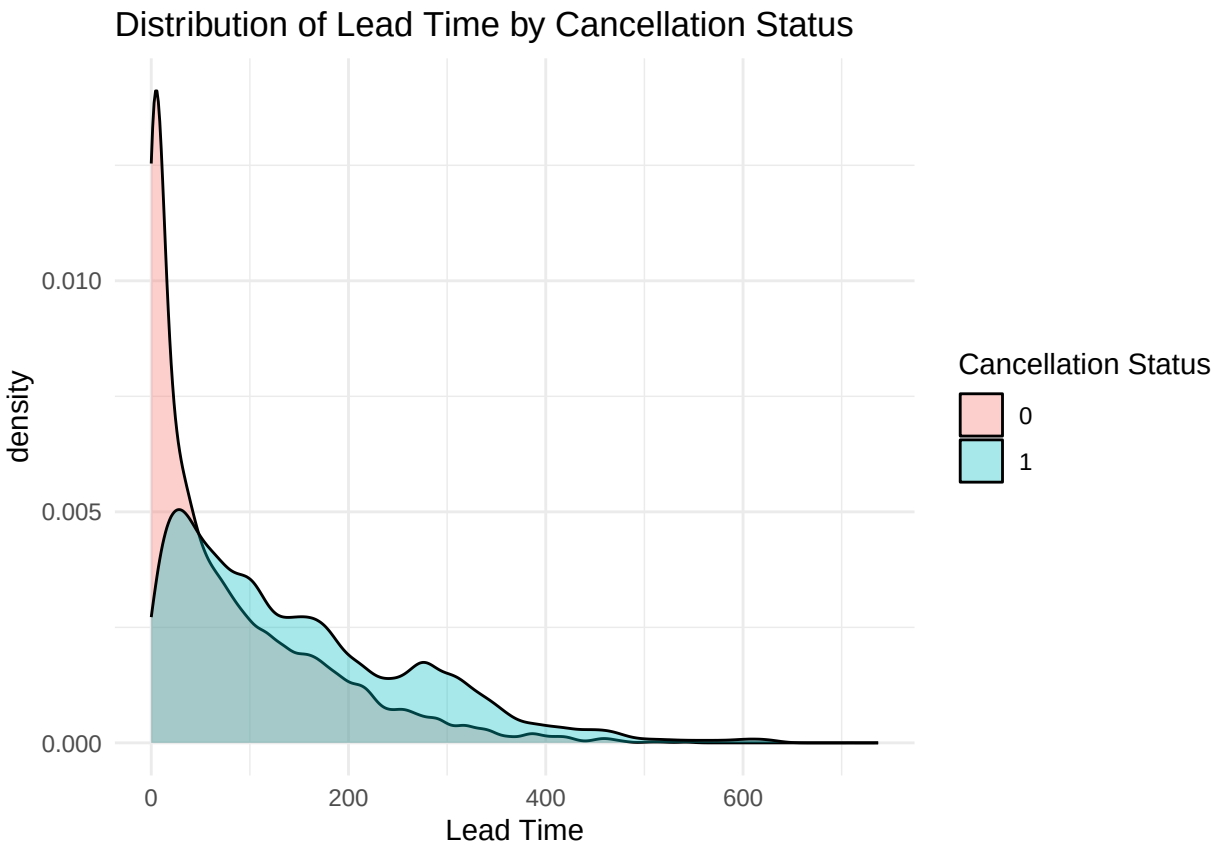
```
## # A tibble: 2 x 7
##   is_canceled     n lt_mean lt_median lt_sd lt_min lt_max
##         <dbl> <int>   <dbl>     <dbl> <dbl>  <dbl>  <dbl>
## 1           0 75166    80.0        45  91.1      0    737
## 2           1 44224   145.        113 119.       0    629
```

Observing these summary statistics alone, we notice that bookings that were canceled tend to have higher lead times than those that were not. We can also visualize this relationship using a boxplot and a density plot:

```
ggplot(hotel_bookings, aes(x = factor(is_canceled), y = lead_time),
       fill=factor(is_canceled)) +
  geom_boxplot() +
  labs(title="Lead Time by Cancellation Status", x="Cancelled (status)", y="Lead Time",
       fill="Cancelled Status") +
  theme_minimal()
```




```r
ggplot(hotel_bookings, aes(x=lead_time, fill=factor(is_canceled))) +
  geom_density(alpha = 0.35) +
  xlim(0, max(hotel_bookings$lead_time)) +
  labs(title="Distribution of Lead Time by Cancellation Status", x="Lead Time",
       fill="Cancellation Status") +
  theme_minimal()
```



This distribution plot suggests that there are more cancelled bookings for which the lead times are greater than there are for honored bookings, as evidenced by the larger density values as lead time increases. This distribution plot also suggests lead time exhibits a strong right-skew (with most of the density at or near a lead time of 0), but we refrain from applying any transformation for our primary model since the exercise dictated we were to only consider transformations prior to formally observing and analyzing our data.

Since we introduce `deposit_type` as a confounder, it makes sense to observe its relationship with `is_canceled` too:

```
table(hotel_bookings$deposit_type, hotel_bookings$is_canceled)
```

```
##
##                   0     1
##   No Deposit  74947 29694
##   Non Refund     93 14494
##   Refundable    126    36
```

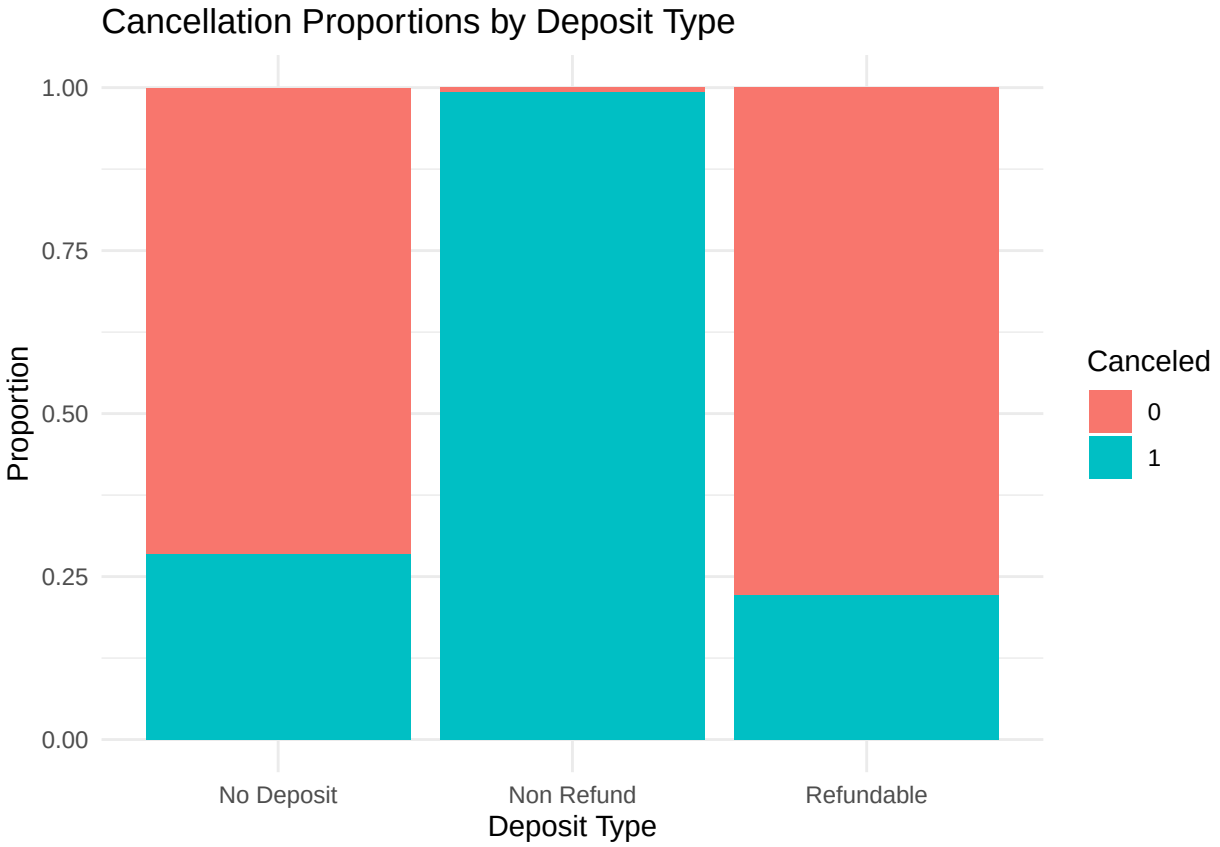
```
prop.table(table(hotel_bookings$deposit_type,
                 hotel_bookings$is_canceled), margin = 1)
```

```
##
##                        0          1
##   No Deposit  0.71622978 0.28377022
##   Non Refund  0.00637554 0.99362446
##   Refundable  0.77777778 0.22222222
```

Interestingly enough, this more detailed breakdown reveals a somewhat shocking conclusion: nearly all of the non-refundable hotel bookings were cancelled. While the explanation as for why may not be immediately obvious, nor might it be obvious how this is associated with `lead_time` (if at all), this still

comes to be quite a surprise. More generally, these summary statistics seem to indicate the relationship between `is_canceled` and `deposit_type` is substantial and highly uneven. We can also understand the relationship graphically:

```
ggplot(hotel_bookings, aes(x = deposit_type, fill = factor(is_canceled))) +
  geom_bar(position = "fill") +
  labs(title = "Cancellation Proportions by Deposit Type", x = "Deposit Type",
    y = "Proportion", fill = "Canceled") +
  theme_minimal()
```



Despite the skewness in `lead_time`, we refrain from making any transformations. The skewness is not too severe where values or relationships might be indiscernible, and, in terms of summary statistics, our standard deviation of 91.1 days (the range is 737 days) does not feel overly large in a way that would warrant the use of a transformation for variance stabilization. And, since one of the goals of this model is its interpretability, we keep `lead_time` as is.

(c) Recall we are testing $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ for our model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot X_i + \beta_2 D_{1i} + \beta_3 D_{2i},$$

where $p_i = \mathbb{P}(Y_i = 1 \mid X_i = x_i)$ denotes the probability of booking i being cancelled given that its observed lead time was x_i days, and where D_{ji} are indicator variables for the non-reference categories of `deposit_type` for each booking i. We fit this model in R using the `glm` function:

```
q2_model <- glm(is_canceled ~ lead_time + deposit_type, data=hotel_bookings,
                family="binomial")
summary(q2_model)
```

```
##
## Call:
## glm(formula = is_canceled ~ lead_time + deposit_type, family = "binomial",
##     data = hotel_bookings)
##
## Coefficients:
##                            Estimate Std. Error  z value Pr(>|z|)
## (Intercept)            -1.253e+00  9.883e-03 -126.756  < 2e-16 ***
## lead_time               3.458e-03  7.078e-05   48.854  < 2e-16 ***
## deposit_typeNon Refund  5.660e+00  1.044e-01   54.230  < 2e-16 ***
## deposit_typeRefundable -5.568e-01  1.911e-01   -2.914  0.00356 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 157398  on 119389  degrees of freedom
## Residual deviance: 123754  on 119386  degrees of freedom
## AIC: 123762
##
## Number of Fisher Scoring iterations: 7
```

The test statistic associated with `lead_time` was 48.854 and its corresponding p-value approximately 0, which means we should reject H_0 at the significance level $\alpha = 0.05$. There is statistically significant evidence to suggest `lead_time` is associated with the probability of a booking being cancelled, after adjusting for the deposit type of the booking. The coefficient for `lead_time` is $\widehat{\beta}_1 \approx 0.00346$, and so the odds ratio is

$$\text{OR} = e^{\widehat{\beta}_1} \approx e^{0.00346} \approx 1.0034.$$

The interpretation of this odds ratio is that holding `deposit_type` constant, each additional one-day increase in lead time is associated with approximately an 0.34% increase in the probability that a booking is to be cancelled. This effect appears to be pretty small on a per-day basis, but it can accumulate over time quite drastically. For instance, a 100-day increase in lead time would correspond to an odds ratio of

$$\text{OR} = e^{\widehat{\beta}_1 \cdot 100} \approx e^{0.346} \approx 1.413.$$

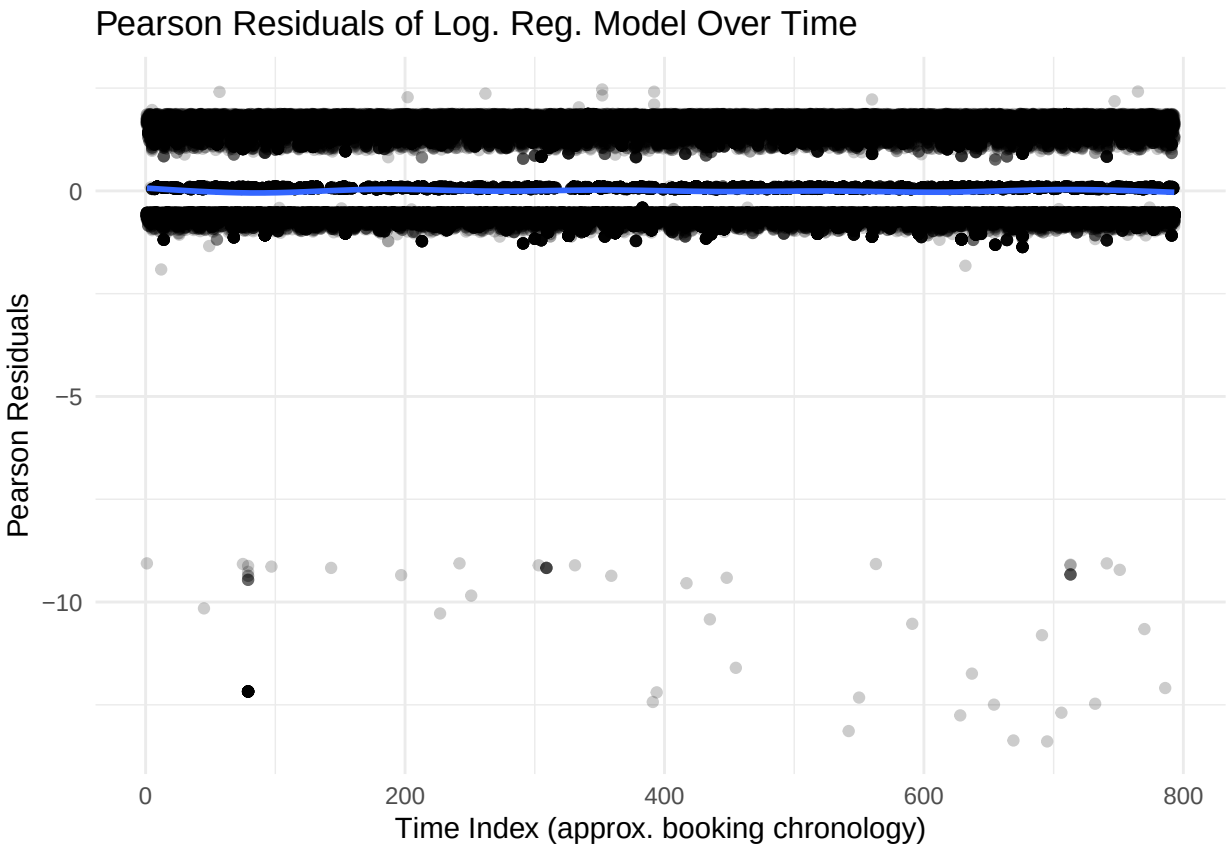
This means that, having fixed the deposit type to be the same across the two bookings, the booking made 100 days earlier is 41% more likely to be cancelled than the later booking.

However, the fact that we have claimed to have found statistically significant evidence does not provide a meaningful or optimal classification threshold for predicting which bookings are going to be cancelled. This because a p-value, like the one we've used to make our decision here, tells us whether a predictor is associated with an outcome on average, not how well a model can empirically separate two classes. Even a statistically significant coefficient from a logistic regression model an yield overlapping predicted probabilities between the cancelled and the non-cancelled bookings. Indeed, choosing the optimal classification threshold is not an inference problem but rather a decision problem.

(d) As part of our secondary analysis, we perform some model diagnostics to determine whether our results can be interpreted as valid. The first condition to check is the binary nature of the outcome variable. This is a trivial check as the condition is satisfied by construction: $Y_i \in$ `is_canceled` $\in \{0, 1\}$. Thus, the logistic regression framework is an appropriate modeling one for this outcome variable.

The second condition to check is the independence of the observations. Each row in the dataset corresponds to a distinct hotel booking, and as there is no explicit structure or repeated-measure identifiers that link more than one row to the same booking event, we can treat the observations to be independent. We check that this treatment via a plot, showcasing how the residuals change as a function of the date of the bookings:

```
hotel_bookings$residuals <- residuals(q2_model, type = "pearson")

hotel_bookings$time_index <- as.numeric(
  interaction(
    hotel_bookings$arrival_date_year,
    hotel_bookings$arrival_date_month,
    hotel_bookings$arrival_date_day_of_month,
    drop = TRUE
  )
)

ggplot(hotel_bookings, aes(x = time_index, y = residuals)) +
  geom_point(alpha = 0.2) +
  geom_smooth(se = FALSE) +
  labs(title = "Pearson Residuals of Log. Reg. Model Over Time",
    x = "Time Index (approx. booking chronology)", y = "Pearson Residuals") +
  theme_minimal()
```

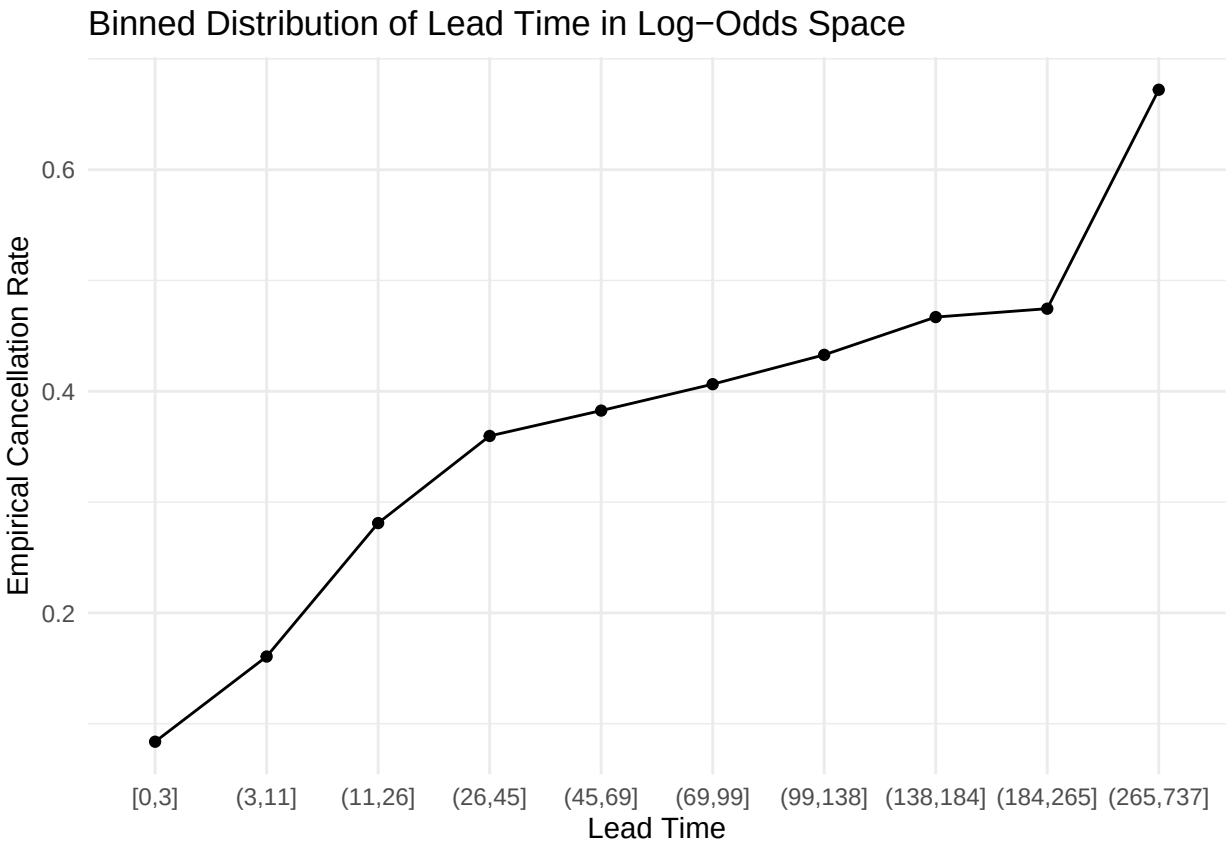


There does not appear to be any systematic curvature or shape to the residuals as plotted here. There are a couple of horizontal clusters of values (indicated by the black "bands"), but these remain roughly the same as the bookings become more and more recent. As there is no strong, visible temporal pattern

in these residuals, we would say that there is no obvious deviation from independence exhibited here.

The last condition to check is linearity, specifically between `lead_time` and the log-odds of cancellation. A standard way to do this is to compare the fitted logits structure against a flexible binned representation of those logits, and check for any noticeable differences. We do this comparison via the following plot:

```
hotel_bookings$lead_bin <- cut(
  hotel_bookings$lead_time,
  breaks = quantile(hotel_bookings$lead_time, probs = seq(0, 1, 0.1), na.rm = TRUE),
  include.lowest = TRUE
)

ggplot(hotel_bookings, aes(x = lead_bin, y = is_canceled)) +
  stat_summary(fun = mean, geom = "point") +
  stat_summary(fun = mean, geom = "line", group = 1) +
  labs(title = "Binned Distribution of Lead Time in Log-Odds Space", x = "Lead Time",
       y = "Empirical Cancellation Rate") +
  theme_minimal()
```



The empirical cancellation rate increases monotonically with lead time. While the increase is not perfectly linear across bins (i.e., between any two bins, the slope of the line is not exactly the same as the slope of the line between the next two bins), there is no strong evidence of some curvature or non-linearity in the cancellation rate which would suggest a violation of the logit-linearity assumption.

Overall, we would say that the conditions for logistic regression are met, and so the results of our primary model are valid. To compare to an alternative, we try to capture some sense for the impact the skewness of `lead_time` by transforming the variable and re-running our analysis from that. That is, instead of modeling raw `lead_time`, we fit a model using `log(lead_time + 1)` as our predictor. This changes the

functional form of our model to

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \log(X_i + 1) + \beta_2 D_{1i} + \beta_3 D_{2i},$$

but does not change that we are testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. There are a few key reasons as to why we might want to apply this particular transformation. First, the lead time values become compressed, meaning the skewness becomes lessened. Second, the compression also helps reduce the variance in the data; with smaller variance, any confidence interval we wish to report for the coefficient estimates becomes more precise. Third, our original model assumes each additional day of lead time yields a constant, additive change in the log-odds of cancellation, and it is entirely possible that increases in lead time actually have diminishing effects at larger values. This transformation helps address this, and the shift by one unit to the right ensures that observations with `lead_time = 0` remain well-defined. The below R code applies this transformation and then fits the updated model:

```
hotel_bookings$log_lead_time <- log(hotel_bookings$lead_time + 1)
q2_alt_model <- glm(is_canceled ~ log_lead_time + deposit_type, data=hotel_bookings,
                    family="binomial")
summary(q2_alt_model)
```

```
##
## Call:
## glm(formula = is_canceled ~ log_lead_time + deposit_type, family = "binomial",
##     data = hotel_bookings)
##
## Coefficients:
##                           Estimate Std. Error  z value Pr(>|z|)
## (Intercept)             -2.271025   0.021782 -104.263   <2e-16 ***
## log_lead_time            0.348230   0.005081   68.534   <2e-16 ***
## deposit_typeNon Refund   5.591029   0.104359   53.575   <2e-16 ***
## deposit_typeRefundable -0.620206   0.191482   -3.239   0.0012 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 157398  on 119389  degrees of freedom
## Residual deviance: 120685  on 119386  degrees of freedom
## AIC: 120693
##
## Number of Fisher Scoring iterations: 7
```

Based on the produced output, the test statistic for `log(lead_time + 1)` is 68.5336613 and its corresponding p-value is roughly 0, which means that we reject the null hypothesis H_0 at the significance level $\alpha = 0.05$. There is statistically significant evidence that longer booking lead times are associated with increased an cancellation probability after adjusting for `deposit_type`, even after having transformed `lead_time`. The estimated coefficient for `log_lead_time` is $\hat{\beta}_1 \approx 0.348$, which is still positive, meaning that any increases in log-transformed lead time will correspond to higher odds of cancellation.

Overall, the substantive conclusions we make as a result of our primary and alternative models are consistent. In both cases, we rejected the null hypothesis and concluded that, holding the deposit type fixed, there exists some positive association between lead time (either transformed or raw) and the probability of cancellation. Moreover, the coefficients for `deposit_type` remain quite stable across both

tests; here, stable is defined as the two estimated coefficient values being close to one another. The fact that our conclusions (and the results themselves, for that matter) are stable and not materially different across the two models suggests that the findings in our primary model are somewhat robust to reasonable changes in model specification.

Despite the general robustness in the findings, there are a couple of limitations to thee models. First, we did not perform a thorough enough investigation to see whether there are any unobserved dependencies between observations in our dataset. These might include repeated bookings from the same customers or agencies, which cannot be fully assessed due to the fact that we were tasked with working with specific variables which might have not been able to showcase those dependencies. Second, and as was the case with **Question 1**, because we made use of an observational dataset, it would not be reasonable to say that the relationships explored in either our primary or alternative models indicate a causal relationship; instead, we must make due with simply saying there exist some associations between the variables.