Report on Expedia Price Sensitivity and Consumer Behavior

Shreya Chandra, Katrin Maliatski, Rimsa Shrestha, Christina Zhu

October 20, 2024

1 Introduction

This report analyzes Expedia's experimental data collected through randomized price variations across several U.S. markets. Namely, the analysis aims to uncover whether a 100 dollar increase in price would lead to a 20 percent decline in booking probability, with the following hypotheses guiding the analysis:

Hypothesis $\mathbf{H_0}$. *A 100 dollar increase in price leads to a 20 percent decrease in booking probability*

Hypothesis $\mathbf{H_1}$. *A 100 dollar increase in price does not lead to a 20 percent decrease in booking probability*

In addition to analyzing how sensitive consumers are to price changes, the analysis aims to uncover differences in price sensitivity by destination and user income. The analysis involves two dependent variables: whether a user booked a hotel (binary outcome) and the number of nights booked (count variable).

2 EDA and Model Selection

EDA was conducted on the Expedia data prior to beginning analysis. In order to analyze user income more effectively, user incomes were grouped into five categories based on quartile distributions: "0-30k", "30-45k", "45-65k", "65-90k", "90k+". The average nights per income group decreased (2.98 → 2.94 nights) as user incomes increased, revealing an unintuitive trend in the data, while price per night and booking rates increased with user income (Figure 1). In terms of destination trends, Las Vegas is most successful in converting customers at a 35 percent booking rate, while Hawaii has the lowest conversion at 20 percent among the four destinations. Average price per night is approximately the same across the destinations (Figure 2).

Additionally, both the dependent variables were plotted against 'Price per Night' to determine which model is most suitable for the analysis (Figures 3 through 6). The plots demonstrated that either linear or logistic regression could be used to model the relationships, therefore, linear regression was used for both 'Booked' and 'Nights' variables.

3 Analysis

A linear regression model was used to analyze the impact of price changes on booking probability. The independent variables include price per night, user income group, and destination. The resulting model was found to be statistically significant (p-value: $<$ 2.2e-16), with a 100 dollar increase in hotel price leading to a decline in booking probability of 8.14 percentage points. This is a much lower effect than hypothesized by Expedia. However, the R-squared value is low (0.09253), indicating that only about 9.25 percent of the variance in booking probability is explained by these variables (Figure 9). To further evaluate whether price sensitivity differs by destination and income, interaction effects were explored between price and these factors. While the model reveals an overall negative effect of price on bookings, the effect varies significantly across regions and income groups (Figure 12). Notably, the highest income group shows decreased price sensitivity, indicating an opportunity for targeted pricing and marketing strategies.

To investigate the impact of price, destination, and user income on the other dependent variable, number of nights, linear regression was also used. This linear regression was run solely on booked data to avoid conflating booking decisions with number of nights. For every 100 dollar increase in price, the booking would decrease by about 0.257 nights. While this model is also significant (p-value $<$ 1.029e-06), the R-squared statistic is extremely low at 0.0055), indicating that only about 0.55 percent of the variance in number of nights stayed probability is explained the variables in the model (Figure 10). An interaction model was generated to further explore the impact of price per night on the number of nights stayed (Figure 13). The model revealed that a higher price per night is associated with a lower number of nights stayed, an association that holds true across destinations and income groups.

4 Conclusion

Our experimental analysis confirms that consumers are sensitive to hotel prices, with a 100 dollar increase in price resulting in an approximate 8.14 percentage points decrease in booking likelihood. This is considerably lower than the 20 percent estimate from observational data. While income does not appear to have a significant impact on booking decisions, there are notable regional differences, with Las Vegas exhibiting higher booking rates than other regions. These insights suggest that pricing strategies should be tailored to individual regions, as sensitivity to price changes may vary across markets. Further experimentation may be needed to refine these estimates and explore additional factors influencing consumer behavior.

401 Homework 1

2024-10-10

```
# Load required libraries
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library('lmtest')
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.4.1
```

```
library(ggplot2)
library(dplyr)
```

```
load("~/Downloads/HW1.Rdata")
colnames(Expedia)[4] <- "Booked"
```

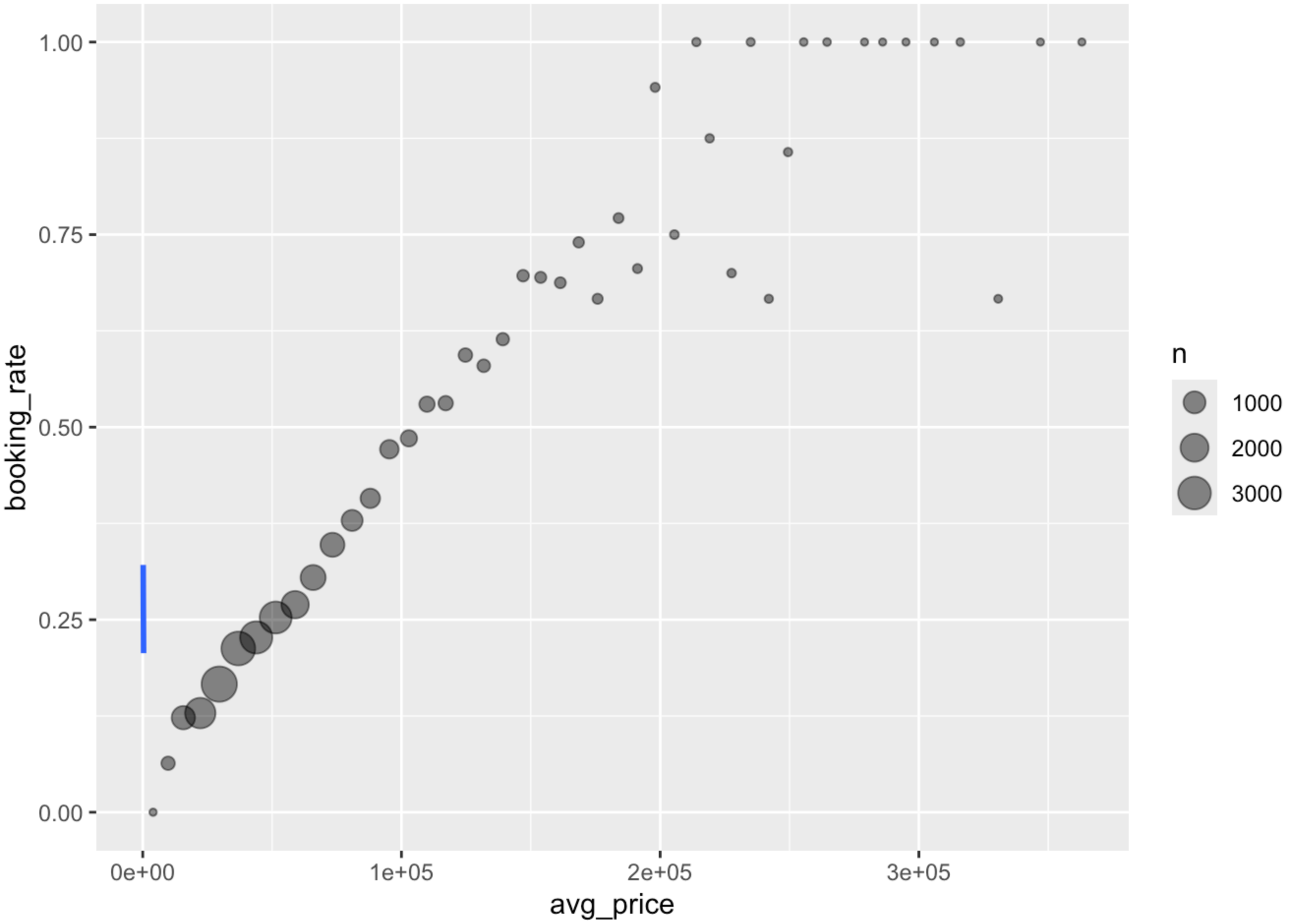
```
summary(Expedia$UserIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4000   31000   45000   52040   65000  363000
```

```
# Group incomes based on quartiles to simplify analysis. Income buckets will allow for a
non-linear shape if one exists. There is a risk of overfitting if there are too many buc
kets. Also, if there are too few buckets, people could be assumed to be similar, howeve
r, in reality they are different. Chosen following buckets ensure equal-sized income buc
kets with each income group having a substantial percentage of the sample.

Expedia$UserIncome_Grouped <- cut(Expedia$UserIncome,
                                   breaks = c(0, 30000, 45000, 65000, 90000, Inf),
                                   labels = c("0-30k", "30-45k", "45-65k", "65-90k", "90
k+"),
                                   include.lowest = TRUE)

# Convert income_group to factor
Expedia$UserIncome_Grouped <- factor(Expedia$UserIncome_Grouped, levels = c("0-30k", "30
-45k", "45-65k", "65-90k", "90k+"))
```

```
## Checking shape of Income with Booked dependent variable. To see whether Income should
be divided into buckets or should be treated as a continuous variable.
binned_data <- Expedia %>%
  mutate(income_bin = cut(UserIncome, breaks = seq(min(UserIncome), max(UserIncome), len
gth.out = 50))) %>%
  group_by(income_bin) %>%
  summarise(
    avg_price = mean(UserIncome),
    booking_rate = mean(Booked),
    n = n()
  )

# Now, let's create the improved plot
ggplot() +
  # Add binned data points
  geom_point(data = binned_data, aes(x = avg_price, y = booking_rate, size = n), alpha =
0.5) +
  geom_smooth(data = Expedia, aes(x = PricePerNight, y = Booked), method = "glm", metho
d.args = list(family = "binomial"), se = FALSE)
```

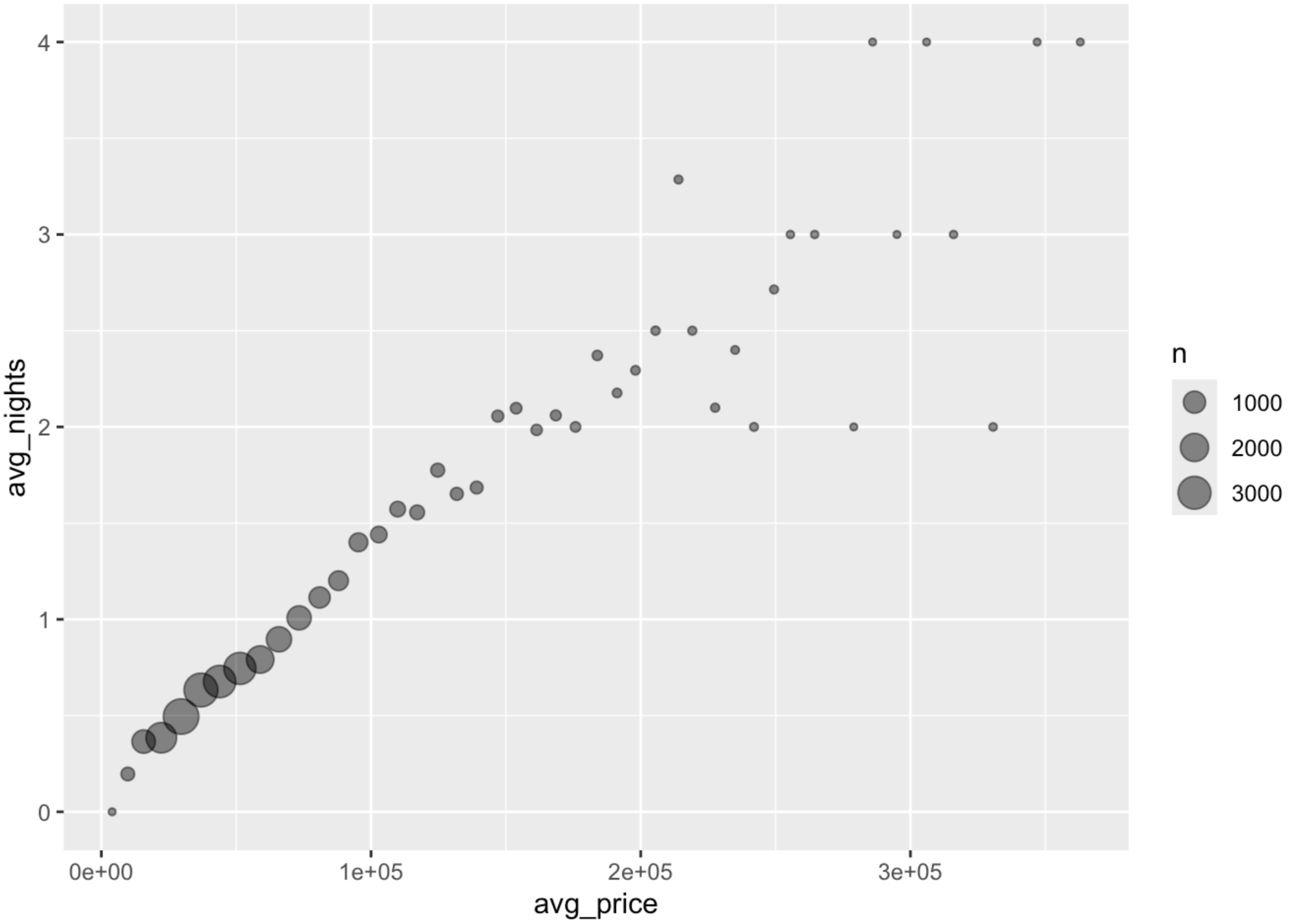
```
## `geom_smooth()` using formula = 'y ~ x'
```



```
## The graph below indicates a non-linear shape. Hence, we divided Income into groups.
```

```
## Checking shape of Income with Nights dependent variable. To see whether Income should
be divided into buckets or should be treated as a continuous variable.
binned_data <- Expedia %>%
  mutate(income_bin = cut(UserIncome, breaks = seq(min(UserIncome), max(UserIncome), len
gth.out = 50))) %>%
  group_by(income_bin) %>%
  summarise(
    avg_price = mean(UserIncome),
    avg_nights = mean(Nights),
    n = n()
  )

# Now, let's create the improved plot
ggplot() +
  # Add binned data points
  geom_point(data = binned_data, aes(x = avg_price, y = avg_nights, size = n), alpha =
0.5) +
  geom_smooth(data = Expedia, aes(x = PricePerNight, y = Nights, color = "Linear"), meth
od = "glm", method.args = list(family = "binomial"), se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Failed to fit group 1.
## Caused by error:
## ! y values must be 0 <= y <= 1
```



```
## The graph below indicates a non-linear shape. Hence, we divided Income into groups.
```

Figure 1

```
booked_data <- Expedia %>%
  filter(Booked == 1)

## Average nights booked by income group
nights_by_income <- booked_data %>%
  group_by(UserIncome_Grouped) %>%
  summarize(avg_nights = mean(Nights)) %>%
  left_join(
    Expedia %>%
      filter(Booked == 1) %>%
      group_by(UserIncome_Grouped) %>%
      summarize(avg_price_per_night_booked = mean(PricePerNight, na.rm = TRUE)),
    by = "UserIncome_Grouped"
  )

print("Average nights booked by income group:")
```

```
## [1] "Average nights booked by income group:"
```

```
print(nights_by_income)
```

```
## # A tibble: 5 × 3
##   UserIncome_Grouped avg_nights avg_price_per_night_booked
##   <fct>                   <dbl>                      <dbl>
## 1 0-30k                    2.98                       246.
## 2 30-45k                   2.98                       247.
## 3 45-65k                   2.94                       248.
## 4 65-90k                   2.93                       247.
## 5 90k+                     2.94                       250.
```

Figure 2

```
## Booking rate by destination
booking_rate_dest <- Expedia %>%
  group_by(UserIncome_Grouped) %>%
  summarize(booking_rate = mean(Booked)) %>%
  left_join(
    Expedia %>%
      filter(Booked == 1) %>%
      group_by(UserIncome_Grouped) %>%
      summarize(avg_price_per_night_booked = mean(PricePerNight, na.rm = TRUE)),
    by = "UserIncome_Grouped"
  )

print("Booking rate by destination:")
```

```
## [1] "Booking rate by destination:"
```

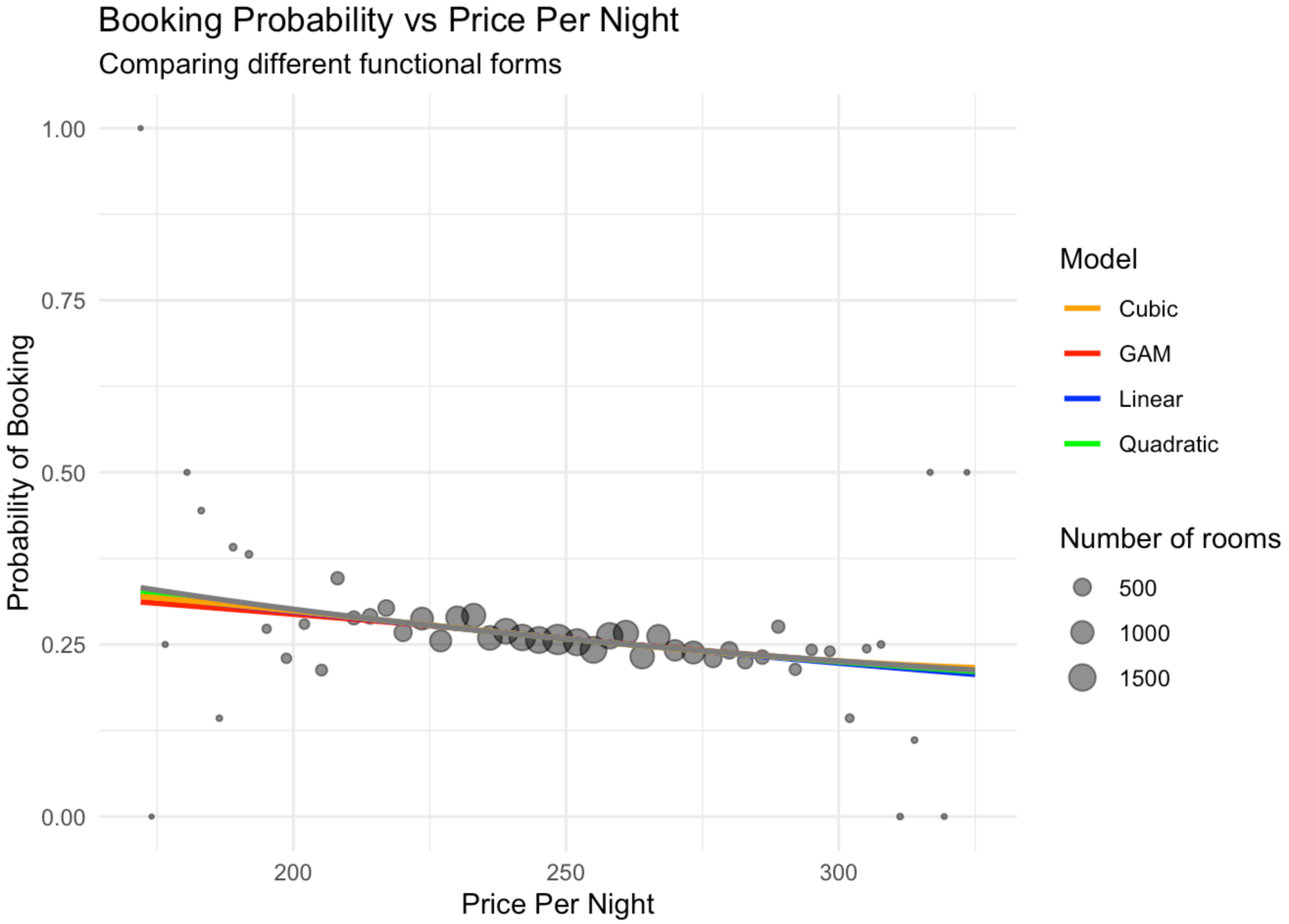
```
print(booking_rate_dest)
```

```
## # A tibble: 5 × 3
##   UserIncome_Grouped booking_rate avg_price_per_night_booked
##   <fct>                     <dbl>                      <dbl>
## 1 0-30k                     0.135                       246.
## 2 30-45k                    0.212                       247.
## 3 45-65k                    0.257                       248.
## 4 65-90k                    0.361                       247.
## 5 90k+                      0.553                       250.
```

Figure 3

```
## Price is the treatment effect variable. However, we will not use the treatment doses approach as Price has too many unique values. Hence, we will not solve using the Non-parametric approach . We will take Price as a continuous treatment.
## create scatterplots to identify what functional form to use for Price Per Night(Booked as dependent variable)
## Since Price per night is a categorical variable, creating scatter plot of mean Price per night and mean Booked dependent variable for Price per night bins.
binned_data <- Expedia %>%
  mutate(price_bin = cut(PricePerNight, breaks = seq(min(PricePerNight), max(PricePerNight), length.out = 50))) %>%
  group_by(price_bin) %>%
  summarise(
    avg_price = mean(PricePerNight),
    booking_rate = mean(Booked),
    n = n()
  )

# scatter plot with binned data
ggplot() +
  geom_point(data = binned_data, aes(x = avg_price, y = booking_rate, size = n), alpha = 0.5) +
  geom_smooth(data = Expedia, aes(x = PricePerNight, y = Booked, color = "GAM"), method = "gam", formula = y ~ s(x, bs = "cs"), se = FALSE) +
  geom_smooth(data = Expedia, aes(x = PricePerNight, y = Booked, color = "Linear"), method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  geom_smooth(data = Expedia, aes(x = PricePerNight, y = Booked, color = "Quadratic"), method = "glm", formula = y ~ poly(x, 2), method.args = list(family = "binomial"), se = FALSE) +
  geom_smooth(data = Expedia, aes(x = PricePerNight, y = Booked, color = "Cubic"), method = "glm", formula = y ~ poly(x, 3), method.args = list(family = "binomial"), se = FALSE) +
  geom_smooth(data = Expedia, aes(x = PricePerNight, y = Booked, color = "Log"), method = "glm", formula = y ~ log(x), method.args = list(family = "binomial"), se = FALSE) +
  # Customize the plot
  scale_color_manual(values = c("GAM" = "red", "Linear" = "blue", "Quadratic" = "green", "Cubic" = "orange")) +
  scale_size_continuous(range = c(0.5, 5), name = "Number of rooms") +
  labs(title = "Booking Probability vs Price Per Night",
       subtitle = "Comparing different functional forms",
       x = "Price Per Night",
       y = "Probability of Booking",
       color = "Model") +
  theme_minimal() +
  theme(legend.position = "right")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



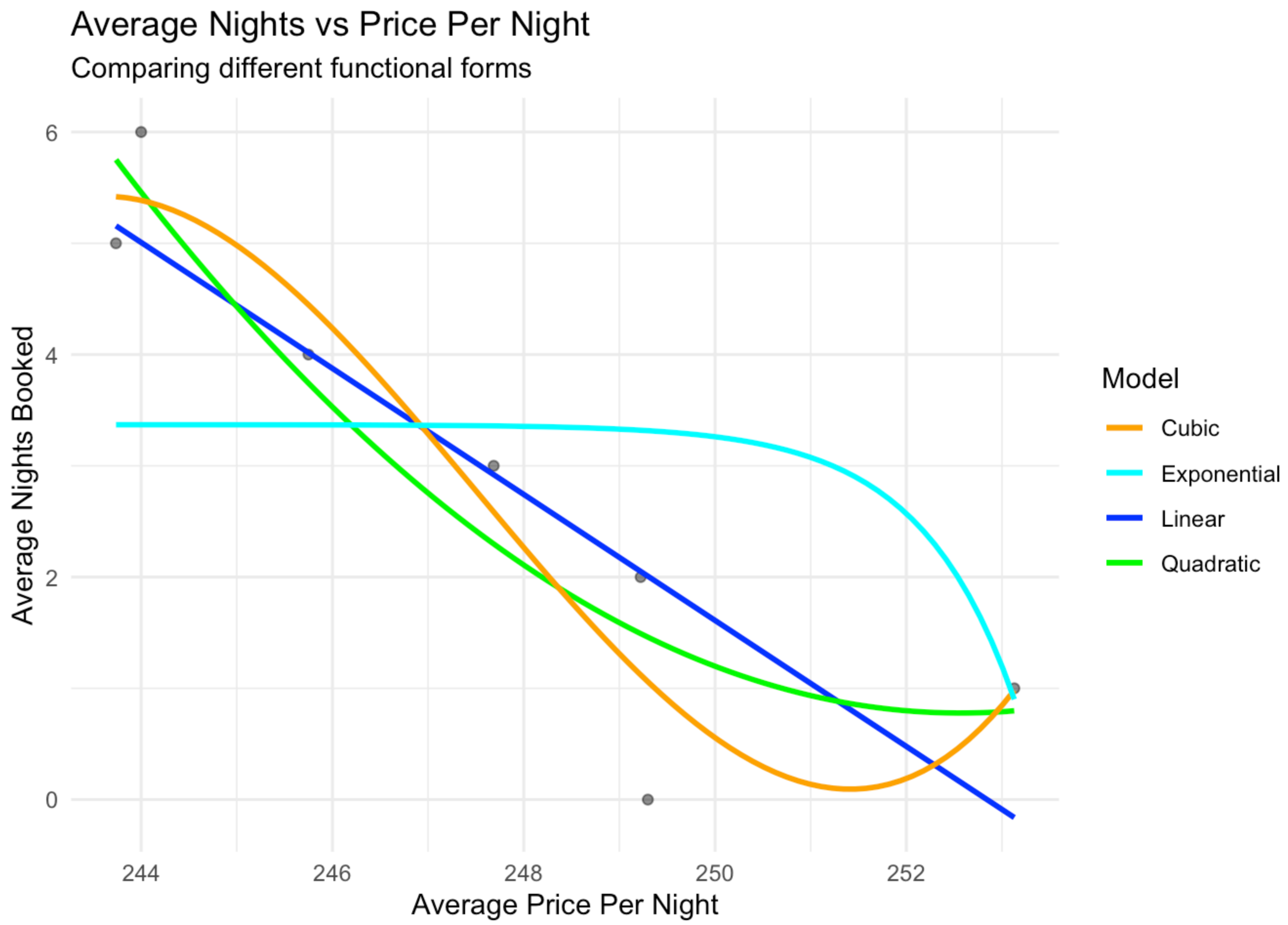
```
##Linear, Quadratic, and Cubic qualify as potential transformations of Price per night
```

Figure 4

```
## Price is the treatment effect variable. However, we will not use the treatment doses approach as Price has too many unique values. Hence, we will not solve using the Non-parametric approach . We will take Price as a continuous treatment.
## create scatterplots to identify what functional form to use for Price Per Night(Nights as dependent variable)
## Since Price per night is a categorical variable, creating scatter plot of mean Price per night and mean Nights dependent variable for Price per night bins.
binned_data <- Expedia %>%
  mutate(nights_bin = cut(Nights, breaks = seq(min(Nights), max(Nights), length.out = 10))) %>%
  group_by(nights_bin) %>%
  summarise(
    avg_price = mean(PricePerNight, na.rm = TRUE),
    avg_nights = mean(Nights, na.rm = TRUE),
    n = n()
  )

# scatter plot with binned data
ggplot() +
  geom_point(data = binned_data, aes(x = avg_price, y = avg_nights), alpha = 0.5) +
  geom_smooth(data = binned_data, aes(x = avg_price, y = avg_nights, color = "GAM"), method = "gam", formula = y ~ s(x, bs = "cs"), se = FALSE) +
  geom_smooth(data = binned_data, aes(x = avg_price, y = avg_nights, color = "Linear"), method = "lm", se = FALSE) +
  geom_smooth(data = binned_data, aes(x = avg_price, y = avg_nights, color = "Quadratic"), method = "lm", formula = y ~ poly(x, 2), se = FALSE) +
  geom_smooth(data = binned_data, aes(x = avg_price, y = avg_nights, color = "Cubic"), method = "lm", formula = y ~ poly(x, 3), se = FALSE) +
  geom_smooth(data = binned_data, aes(x = avg_price, y = avg_nights, color = "Exponential"), method = "glm", formula = y ~ exp(x), method.args = list(family = "gaussian"), se = FALSE) +
  # Customize the plot
  scale_color_manual(values = c("GAM" = "red", "Linear" = "blue", "Quadratic" = "green", "Cubic" = "orange", "Exponential" = "cyan")) +
  labs(title = "Average Nights vs Price Per Night",
       subtitle = "Comparing different functional forms",
       x = "Average Price Per Night",
       y = "Average Nights Booked",
       color = "Model") +
  theme_minimal() +
  theme(legend.position = "right")
```

```
## Warning: Failed to fit group 1.
## Caused by error in `smooth.construct.cr.smooth.spec()`:
## ! x has insufficient unique values to support 10 knots: reduce k.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
##Linear, Quadratic and Cubic qualify as potential transformations of Price per night
```

Figure 5

```
ggplot(Expedia, aes(x = PricePerNight, y = Booked, color = Region)) +
  stat_summary(fun = mean, geom = "point", size = 2) +
  geom_smooth(method = "loess", se = FALSE, size = 1) +
  scale_y_continuous(breaks = c(0, 1), labels = c("Not Booked", "Booked")) +
  scale_x_continuous(breaks = seq(0, max(Expedia$PricePerNight), by = 50)) +
  labs(title = "Average Booking Rate by Price and City",
       x = "Price Per Night ($)",
       y = "Booking Status",
       color = "City") +
  theme_minimal() +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

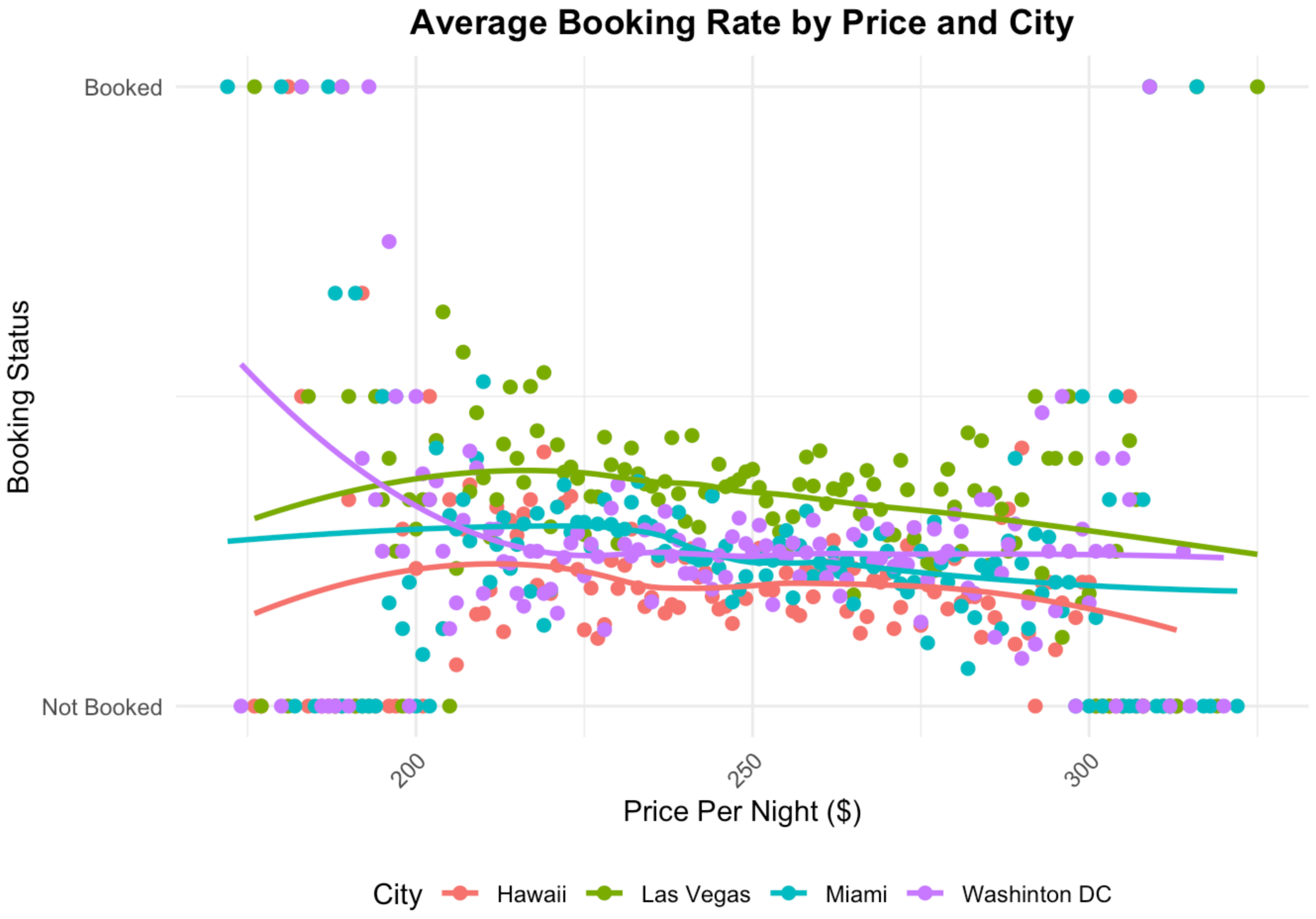
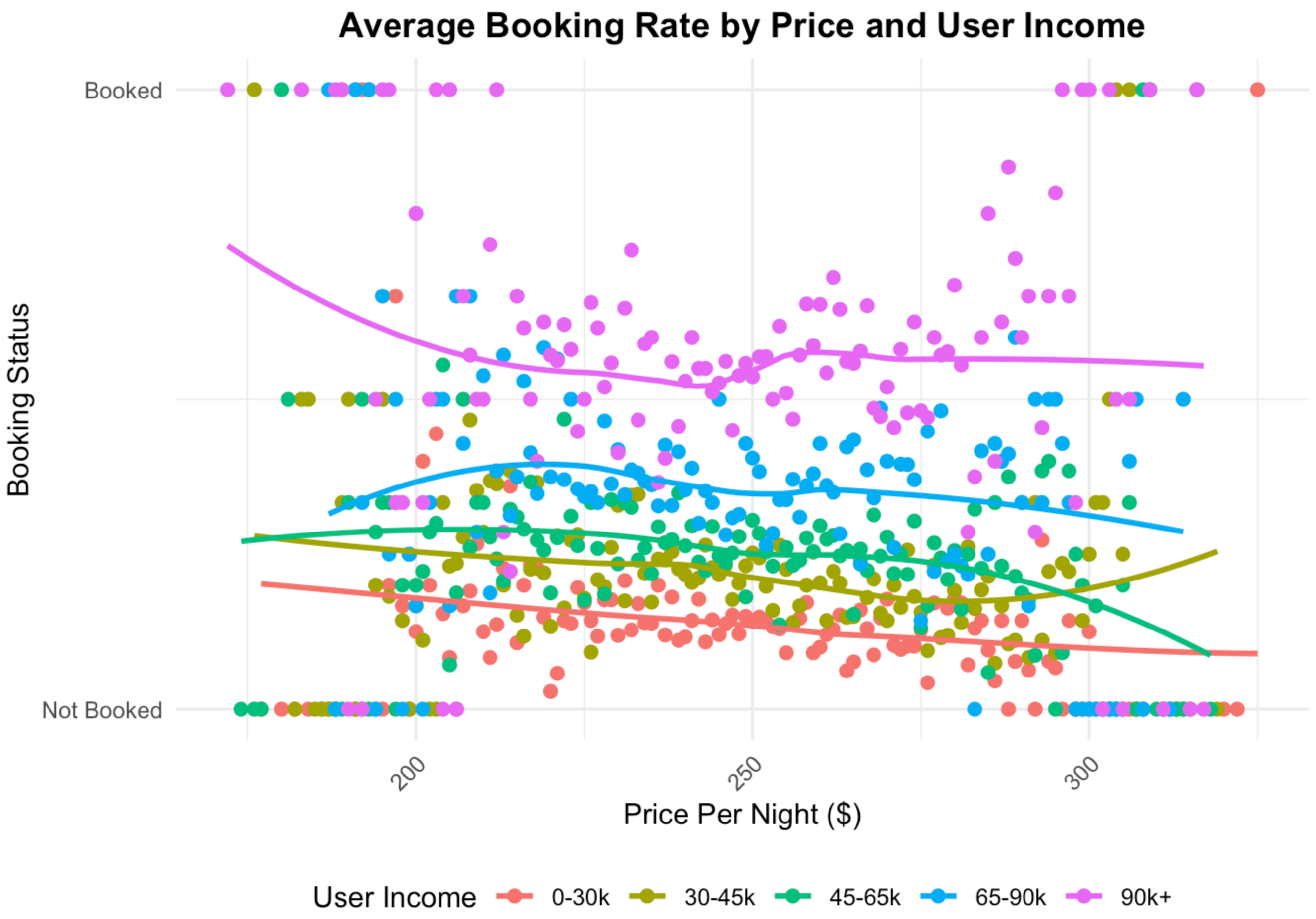


Figure 6

```
ggplot(Expedia, aes(x = PricePerNight, y = Booked, color = UserIncome_Grouped)) +
  stat_summary(fun = mean, geom = "point", size = 2) +
  geom_smooth(method = "loess", se = FALSE, size = 1) +
  scale_y_continuous(breaks = c(0, 1), labels = c("Not Booked", "Booked")) +
  scale_x_continuous(breaks = seq(0, max(Expedia$PricePerNight), by = 50)) +
  labs(title = "Average Booking Rate by Price and User Income",
       x = "Price Per Night ($)",
       y = "Booking Status",
       color = "User Income") +
  theme_minimal() +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



#

Figure 7

```
hist(Expedia$UserIncome,
     main = "Distribution of User Incomes",
     xlab = "User Income",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")
```

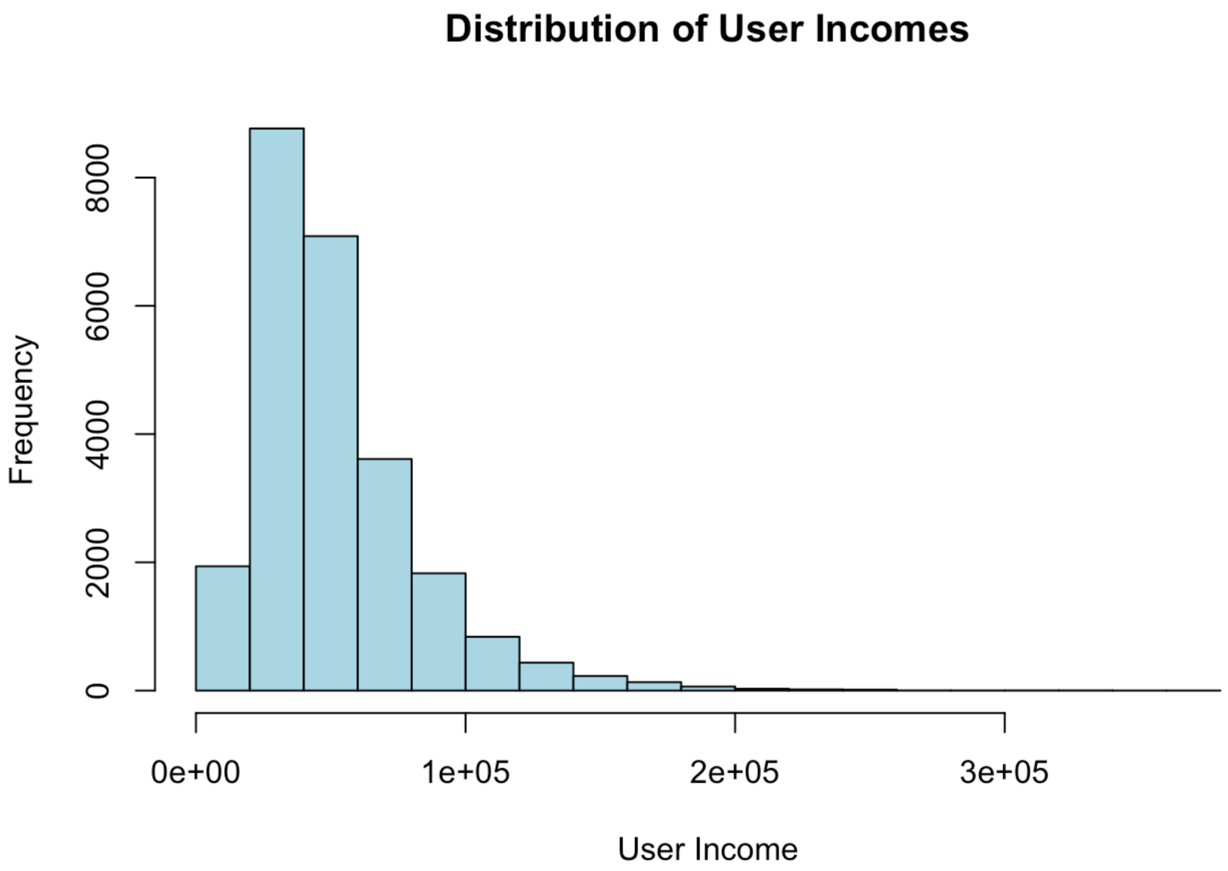
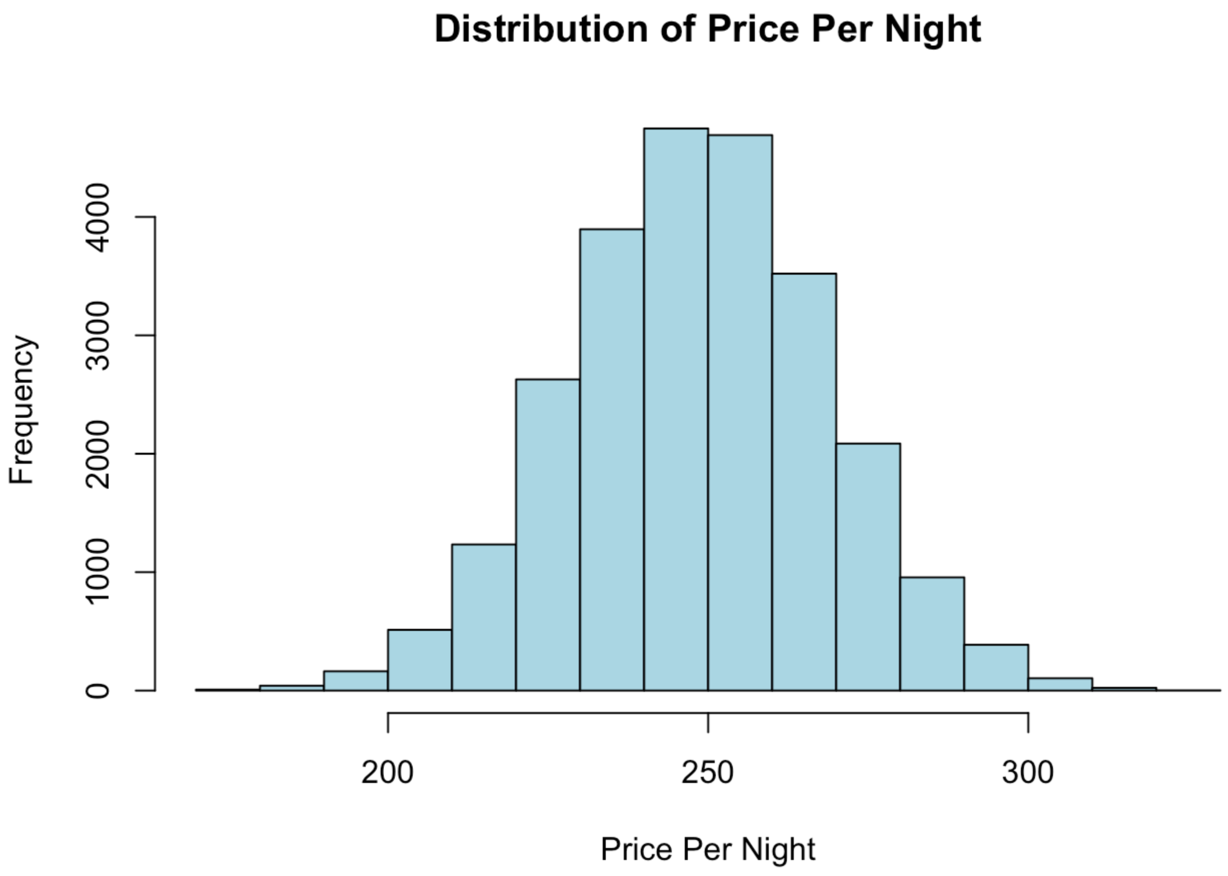


Figure 8

```
hist(Expedia$PricePerNight,
     main = "Distribution of Price Per Night",
     xlab = "Price Per Night",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")
```



```
## Correlation between price and booking
cor_price_booking <- cor(Expedia$PricePerNight, Expedia$Booked)
print(paste("Correlation between price and booking:", cor_price_booking))
```

```
## [1] "Correlation between price and booking: -0.0345819092624071"
```

Figure 9

```
## Modeling
# Linear regression for booking probability
model_booking <- lm(Booked ~ PricePerNight + Region + UserIncome,
                     data = Expedia)

## Booking probability model
summary(model_booking)
```

```
## 
## Call:
## lm(formula = Booked ~ PricePerNight + Region + UserIncome, data = Expedia)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.3738 -0.2647 -0.1709  0.3166  0.9561 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         1.915e-01  3.307e-02   5.791 7.10e-09 ***
## PricePerNight      -8.138e-04  1.303e-04  -6.246 4.28e-10 ***
## RegionLas Vegas     1.488e-01  7.453e-03  19.965  < 2e-16 ***
## RegionMiami         4.635e-02  7.453e-03   6.219 5.10e-10 ***
## RegionWashinton DC  4.613e-02  7.453e-03   6.189 6.15e-10 ***
## UserIncome          4.045e-06  8.512e-08  47.517  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4166 on 24994 degrees of freedom
## Multiple R-squared:  0.09758,    Adjusted R-squared:  0.0974 
## F-statistic: 540.5 on 5 and 24994 DF,  p-value: < 2.2e-16
```

Figure 10

```
# Linear regression for number of nights
model_nights <- lm(Nights ~ PricePerNight + Region + UserIncome, data = booked_data)

## Number of nights model
summary(model_nights)
```

```
##
## Call:
## lm(formula = Nights ~ PricePerNight + Region + UserIncome, data = booked_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10532 -0.85583  0.03969  0.13551  3.02842
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.593e+00  1.170e-01  30.714  < 2e-16 ***
## PricePerNight      -2.569e-03  4.638e-04  -5.539 3.16e-08 ***
## RegionLas Vegas     3.299e-02  2.693e-02   1.225    0.221
## RegionMiami        -9.923e-05  2.874e-02  -0.003    0.997
## RegionWashinton DC -1.025e-02  2.877e-02  -0.356    0.722
## UserIncome         -1.767e-07  2.390e-07  -0.739    0.460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7546 on 6488 degrees of freedom
## Multiple R-squared:  0.005505,   Adjusted R-squared:  0.004739
## F-statistic: 7.183 on 5 and 6488 DF,  p-value: 1.029e-06
```

Figure 11

```
# Price sensitivity calculation
avg_price <- mean(Expedia$PricePerNight)
price_coef <- coef(model_booking)["PricePerNight"]
price_sensitivity <- (exp(price_coef * (avg_price + 100)) - exp(price_coef * avg_price))
/ exp(price_coef * avg_price)
print(paste("Estimated decrease in booking probability for $100 price increase:",
            round(price_sensitivity * 100, 2), "%"))
```

```
## [1] "Estimated decrease in booking probability for $100 price increase: -7.82 %"
## Since the three potential transformations of Price per night variable were Linear, Quadratic and Cubic.
## For dependent variable Booked or not, we built Interaction models with all four transformations one-by-one.
x <- Expedia$PricePerNight
PricePerNight <- x
model_booking_interact2 <- lm(Booked ~ PricePerNight * Region + PricePerNight * UserIncome_Grouped,
                              data = Expedia)
model_summary <- summary(model_booking_interact2)
# Extract just the adjusted R-squared
adjusted_r1 <- model_summary$adj.r.squared
# Print the adjusted R-squared value
sprintf("Adjusted R^2 for Linear model is: %.6f ", adjusted_r1)
```

```
## [1] "Adjusted R^2 for Linear model is: 0.091986 "
```

```
x <- Expedia$PricePerNight
PricePerNightQuad <- x + I(x^2)
model_booking_interact2 <- lm(Booked ~ PricePerNightQuad * Region + PricePerNightQuad * UserIncome_Grouped,
                              data = Expedia)
model_summary <- summary(model_booking_interact2)
# Extract just the adjusted R-squared
adjusted_r2 <- model_summary$adj.r.squared
# Print the adjusted R-squared value
sprintf("Adjusted R^2 for Quadratic model is: %.6f ", adjusted_r2)
```

```
## [1] "Adjusted R^2 for Quadratic model is: 0.092012 "
```

```
x <- Expedia$PricePerNight
PricePerNightCubic <- x + I(x^2) + I(x^3)
model_booking_interact3 <- lm(Booked ~ PricePerNightCubic * Region + PricePerNightCubic * UserIncome_Grouped,
                              data = Expedia)
model_summary <- summary(model_booking_interact3)
# Extract just the adjusted R-squared
adjusted_r3 <- model_summary$adj.r.squared
# Print the adjusted R-squared value
sprintf("Adjusted R^2 for Cubic model is: %.6f ", adjusted_r3)
```

```
## [1] "Adjusted R^2 for Cubic model is: 0.092021 "
```

```
## We checked adjusted R^2 for each transformation of Price per night and did not see much improvement with any of the transformations. Hence, we kept Price per night without any transformation in the final interaction model.
```

Figure 12

```
# Interaction models
## We choose a linear regression model as compared to a logistic regression model as gra
ph of Booked Vs each of the independent variables (Price treatment, Income covariate, an
d Regioc covariate) showed a linear trend. Hence, despite the fact that Booked is a cate
gorical variable, we see a linear trend in our sample data in the region of interest. He
nce, we use a linear model. Also, for the purpose of this assignment, we do not forecast
values of Booked. Hence, a linear model serves our purpose.
## Since working with multiple interactions is challenging, we use second approach of in
cluding treatment variable in our model, that is including treatment variable by itself
as an independent variable
## We don’t interact the two covariate variables with each-other
model_booking_interact <- lm(Booked ~ PricePerNight * Region + PricePerNight * UserIncom
e_Grouped,
                              data = Expedia)
summary(model_booking_interact)
```

```
## 
## Call:
## lm(formula = Booked ~ PricePerNight * Region + PricePerNight * 
##     UserIncome_Grouped, data = Expedia)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.6439 -0.2533 -0.1783  0.3603  0.9502 
## 
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             2.353e-01  8.816e-02   2.669  0.00762
## PricePerNight                          -6.463e-04  3.535e-04  -1.828  0.06754
## RegionLas Vegas                         2.832e-01  9.220e-02   3.071  0.00213
## RegionMiami                             2.127e-01  9.231e-02   2.305  0.02120
## RegionWashinton DC                     -2.957e-02  9.215e-02  -0.321  0.74831
## UserIncome_Grouped30-45k                1.158e-01  9.172e-02   1.263  0.20665
## UserIncome_Grouped45-65k                1.497e-01  9.382e-02   1.596  0.11059
## UserIncome_Grouped65-90k                2.244e-01  1.097e-01   2.045  0.04088
## UserIncome_Grouped90k+                  1.064e-01  1.219e-01   0.873  0.38277
## PricePerNight:RegionLas Vegas          -5.444e-04  3.695e-04  -1.473  0.14066
## PricePerNight:RegionMiami              -6.686e-04  3.699e-04  -1.808  0.07069
## PricePerNight:RegionWashinton DC        3.026e-04  3.691e-04   0.820  0.41228
## PricePerNight:UserIncome_Grouped30-45k -1.536e-04  3.674e-04  -0.418  0.67590
## PricePerNight:UserIncome_Grouped45-65k -1.040e-04  3.757e-04  -0.277  0.78193
## PricePerNight:UserIncome_Grouped65-90k  4.852e-06  4.404e-04   0.011  0.99121
## PricePerNight:UserIncome_Grouped90k+    1.252e-03  4.868e-04   2.572  0.01012
##                                           
## (Intercept)                            ** 
## PricePerNight                          .  
## RegionLas Vegas                        ** 
## RegionMiami                            *  
## RegionWashinton DC                        
## UserIncome_Grouped30-45k                  
## UserIncome_Grouped45-65k                  
## UserIncome_Grouped65-90k               *  
## UserIncome_Grouped90k+                    
## PricePerNight:RegionLas Vegas             
## PricePerNight:RegionMiami              .  
## PricePerNight:RegionWashinton DC          
## PricePerNight:UserIncome_Grouped30-45k    
## PricePerNight:UserIncome_Grouped45-65k    
## PricePerNight:UserIncome_Grouped65-90k    
## PricePerNight:UserIncome_Grouped90k+   *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4179 on 24984 degrees of freedom
## Multiple R-squared:  0.09253,    Adjusted R-squared:  0.09199 
## F-statistic: 169.8 on 15 and 24984 DF,  p-value: < 2.2e-16
## Since the four potential transformations of Price per night variable were Linear, Quadratic and Cubic.
## For dependent variable Number of nights booked, we built Interaction models with all four transformations one-by-one.
x <- Expedia$PricePerNight
PricePerNight <- x
model_booking_interact2 <- lm(Nights ~ PricePerNight * Region + PricePerNight * UserIncome_Grouped,
                              data = Expedia)
model_summary <- summary(model_booking_interact2)
# Extract just the adjusted R-squared
adjusted_r1 <- model_summary$adj.r.squared
# Print the adjusted R-squared value
sprintf("Adjusted R^2 for Linear model is: %.6f ", adjusted_r1)
```

```
## [1] "Adjusted R^2 for Linear model is: 0.084251 "
```

```
x <- Expedia$PricePerNight
PricePerNightQuad <- x + I(x^2)
model_booking_interact2 <- lm(Nights ~ PricePerNightQuad * Region + PricePerNightQuad * UserIncome_Grouped,
                              data = Expedia)
model_summary <- summary(model_booking_interact2)
# Extract just the adjusted R-squared
adjusted_r2 <- model_summary$adj.r.squared
# Print the adjusted R-squared value
sprintf("Adjusted R^2 for Quadratic model is: %.6f ", adjusted_r2)
```

```
## [1] "Adjusted R^2 for Quadratic model is: 0.084312 "
```

```
x <- Expedia$PricePerNight
PricePerNightCubic <- x + I(x^2) + I(x^3)
model_booking_interact3 <- lm(Nights ~ PricePerNightCubic * Region + PricePerNightCubic * UserIncome_Grouped,
                              data = Expedia)
model_summary <- summary(model_booking_interact3)
# Extract just the adjusted R-squared
adjusted_r3 <- model_summary$adj.r.squared
# Print the adjusted R-squared value
sprintf("Adjusted R^2 for Cubic model is: %.6f ", adjusted_r3)
```

```
## [1] "Adjusted R^2 for Cubic model is: 0.084351 "
```

```
## We checked adjusted R^2 for each transformation of Price per night and did not see much improvement with any of the transformations. Hence, we kept Price per night without any transformation in the final interaction model.
```

Figure 13

```
## We choose a linear regression model as compared to a logistic regression model as gra
ph of Nights Vs each of the independent variables (Price treatment, Income covariate, an
d Regioc covariate) showed a linear trend. Hence, despite the fact that Nights is a cate
gorical variable, we see a linear trend in our sample data in the region of interest. He
nce, we use a linear model. Also, for the purpose of this assignment, we do not forecast
values of Nights. Hence, a linear model serves our purpose.
## Since working with multiple interactions is challenging, we use second approach of in
cluding treatment variable in our model, that is including treatment variable by itself
as an independent variable
 ## We don’t interact the two covariate variables with each-other
model_nights_interact <- lm(Nights ~ PricePerNight * Region + PricePerNight * UserIncome
_Grouped,
                            data = Expedia)
summary(model_nights_interact)
```

```
##
## Call:
## lm(formula = Nights ~ PricePerNight * Region + PricePerNight *
##     UserIncome_Grouped, data = Expedia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0040 -0.7506 -0.5244  0.4131  4.9277
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              0.7233594  0.2728290   2.651  0.00802
## PricePerNight                           -0.0020158  0.0010941  -1.842  0.06544
## RegionLas Vegas                          0.9036632  0.2853345   3.167  0.00154
## RegionMiami                              0.6468765  0.2856918   2.264  0.02357
## RegionWashinton DC                      -0.1671595  0.2851777  -0.586  0.55777
## UserIncome_Grouped30-45k                 0.4845142  0.2838581   1.707  0.08786
## UserIncome_Grouped45-65k                 0.5125315  0.2903626   1.765  0.07755
## UserIncome_Grouped65-90k                 0.9274210  0.3395833   2.731  0.00632
## UserIncome_Grouped90k+                   0.7503902  0.3771228   1.990  0.04663
## PricePerNight:RegionLas Vegas           -0.0018353  0.0011434  -1.605  0.10848
## PricePerNight:RegionMiami               -0.0020481  0.0011447  -1.789  0.07359
## PricePerNight:RegionWashinton DC         0.0012014  0.0011422   1.052  0.29292
## PricePerNight:UserIncome_Grouped30-45k -0.0010166  0.0011372  -0.894  0.37135
## PricePerNight:UserIncome_Grouped45-65k -0.0006144  0.0011627  -0.528  0.59719
## PricePerNight:UserIncome_Grouped65-90k -0.0011048  0.0013629  -0.811  0.41759
## PricePerNight:UserIncome_Grouped90k+    0.0019061  0.0015065   1.265  0.20580
##
## (Intercept)                            **
## PricePerNight                          .
## RegionLas Vegas                        **
## RegionMiami                            *
## RegionWashinton DC
## UserIncome_Grouped30-45k               .
## UserIncome_Grouped45-65k               .
## UserIncome_Grouped65-90k               **
## UserIncome_Grouped90k+                 *
## PricePerNight:RegionLas Vegas
## PricePerNight:RegionMiami              .
## PricePerNight:RegionWashinton DC
## PricePerNight:UserIncome_Grouped30-45k
## PricePerNight:UserIncome_Grouped45-65k
## PricePerNight:UserIncome_Grouped65-90k
## PricePerNight:UserIncome_Grouped90k+
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.293 on 24984 degrees of freedom
## Multiple R-squared:  0.0848, Adjusted R-squared:  0.08425
## F-statistic: 154.3 on 15 and 24984 DF,  p-value: < 2.2e-16
```