BAX 401 Case 1 Expedia Homework 1

Tushar Yadav - Vatsal Nanawati - Wenjun Song - April Yang

Section 1

Model Free Analysis Insights

Distribution of Price Per Night: The histogram *titled "Distribution of Price Per Night (Expedia)"* of the Price Per Night distribution shows that most of the hotels charge between $200 and $300 with a median as $250 having the maximum frequency.

We may also conclude that the pricing approach is primarily aimed at customers who are budget conscious, with prices starting about $250 per night, however premium offerings are available for higher-paying customers and cheaper hotels for those on a tight budget(we can notice the skewness on both end if we look at the range $200 to $300).

Price Sensitivity by Region: The scatterplot *titled "Average Booking Probability by Price and City"* for Price Per Night vs Average Booking Probability by Region indicates that booking probability decreases as price increases across all regions. We can see that the average is towards the lower end of the 0 and 1 spectrum with a negative slope. Although 0 and 1 does not hold true to their values as they represent status of "Not Booked" and "Booked" respectively.

Regions with higher-priced hotels have a sharper fall in booking, whereas regions with lower average costs exhibit a more mild sensitivity to price fluctuations. This suggests that customers in specific locations are more price-sensitive, and regional pricing tactics should be tailored to maximize bookings.

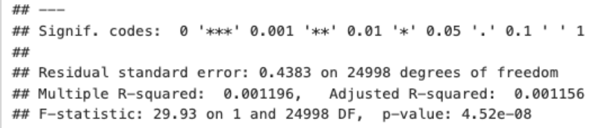
Price Sensitivity by User Income: The scatterplot *titled "Average Booking Probability by Price and User Income"* of Price Per Night vs Average Booking by User Income shows higher-income customers are less price sensitive than lower-income customers. Despite rising prices, the booking likelihood for higher-income individuals stays rather consistent.

Lower-income customers experience a faster fall in booking probability as prices rise, showing that they are more sensitive to price fluctuations. This shows that income level influences price sensitivity, and personalized pricing tactics should target different income bands to increase bookings.

Simple Linear Regression Model Analysis Insights

Booked vs. Price Per Night:

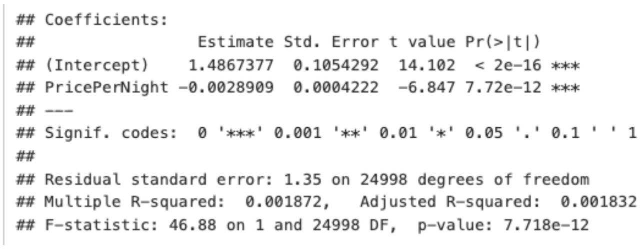
The simple linear regression model's R-squared value of 0.001196 indicates that price per night alone explains only 0.12% of the variation in whether a consumer books a hotel and about 7.4% decrease for $100 increase in price. This shows that price is not the only factor influencing booking behavior, and that other factors like hotel quality, customer preferences, and availability are likely to have a greater impact in a realistic case.



```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.4463765  0.0342230  13.043  < 2e-16 ***
## PricePerNight -0.0007498  0.0001371  -5.471 4.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4383 on 24998 degrees of freedom
## Multiple R-squared:  0.001196,   Adjusted R-squared:  0.001156
## F-statistic: 29.93 on 1 and 24998 DF,  p-value: 4.52e-08
```

Nights vs. Price Per Night:

Similarly, the regression of Nights Stayed vs. Price Per Night yields an R-squared value of 0.001872, indicating that price has little influence on how long customers stay at a hotel and about 28.9% decrease for $100 increase in price. This suggests that the number of nights booked is more likely determined by other factors, such as the purpose of the trip, than by the hotel's price.

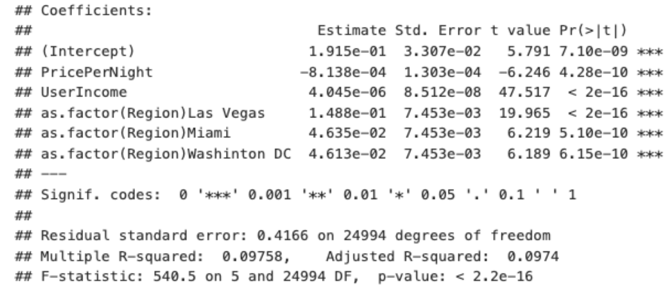


```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.4867377  0.1054292  14.102  < 2e-16 ***
## PricePerNight -0.0028909  0.0004222  -6.847 7.72e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.35 on 24998 degrees of freedom
## Multiple R-squared:  0.001872,   Adjusted R-squared:  0.001832
## F-statistic: 46.88 on 1 and 24998 DF,  p-value: 7.718e-12
```

Multiple Linear Regression Model Analysis Insights

Booked vs. Price Per Night, Income, and Region:

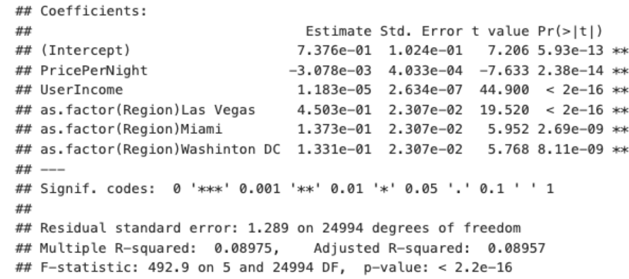
The multiple linear regression model (R-squared = 0.09758) reveals that Price Per Night, Income, and Region combined account for approximately 9.76% of the variation in booking behavior. While the R-squared value remains very low, it demonstrates that income and geography have additional explanatory power, implying that particular income levels and regions are more sensitive to price changes than others. This shows that pricing strategies should take into account both customer income and region to better predict booking outcomes. This suggests that multivariable models are more effective at predicting booking behavior than price alone.



```
## 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 ***
## UserIncome                   4.045e-06  8.512e-08  47.517  < 2e-16 ***
## as.factor(Region)Las Vegas   1.488e-01  7.453e-03  19.965  < 2e-16 ***
## as.factor(Region)Miami       4.635e-02  7.453e-03   6.219 5.10e-10 ***
## as.factor(Region)Washinton DC  4.613e-02  7.453e-03   6.189 6.15e-10 ***
## ---
## 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
```

Nights vs. Price Per Night, Income, and Region:

In the second multiple regression model (R-squared = 0.08975), Price Per Night, Income, and Region account for approximately 8.98% of the variation in the number of nights booked. While not very high, this model indicates that income and region have a moderate impact on length of stay, however other factors are likely to play a larger role. This model indicates that length of stay decisions are likely more influenced by personal factors rather than just price, suggesting that Expedia could benefit from understanding customer trip motivations and providing targeted offers for longer stays.



```
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  7.376e-01  1.024e-01   7.206 5.93e-13 ***
## PricePerNight               -3.078e-03  4.033e-04  -7.633 2.38e-14 ***
## UserIncome                   1.183e-05  2.634e-07  44.900  < 2e-16 ***
## as.factor(Region)Las Vegas   4.503e-01  2.307e-02  19.520  < 2e-16 ***
## as.factor(Region)Miami       1.373e-01  2.307e-02   5.952 2.69e-09 ***
## as.factor(Region)Washinton DC  1.331e-01  2.307e-02   5.768 8.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.289 on 24994 degrees of freedom
## Multiple R-squared:  0.08975,    Adjusted R-squared:  0.08957
## F-statistic: 492.9 on 5 and 24994 DF,  p-value: < 2.2e-16
```

Comparing the Model Free Analysis and Model Analysis

The prior estimate using model free analysis suggested that a $100 increase in price would lead to approximately 20% decrease in the likelihood of booking. The experimental results from the linear regression models suggest a much weaker relationship between price and booking behavior, indicating that the observational estimate may have overstated the impact of price increases.

The inclusion of income and region in the multiple regression models slightly improves the explanatory power, suggesting that price sensitivity is moderated by these factors, but the overall impact of price on booking behavior is less substantial than initially thought.

Final Conclusions

When analyzed in distinct ways, price sensitivity is limited, with both simple and multiple regression models providing minimal statistical significance. However, income and area add complexities to price sensitivity, with lower-income clients and specific regions being more sensitive to price fluctuations.

The observational estimate of a 20% decline in bookings for every $100 rise in price appears to be overestimated, and experimental data suggests that other factors play a larger role in booking decisions. Targeted pricing tactics based on customer income and destination region may assist optimize bookings, but price alone is not a powerful lever for influencing customer behavior in this setting.

iii_expedia

2024-10-14

Homework Case 1

Tushar Yadav - Vatsal Nanawati - Wenjun Song - April Yang

Section 1

```
install.packages('ggplot2')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

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

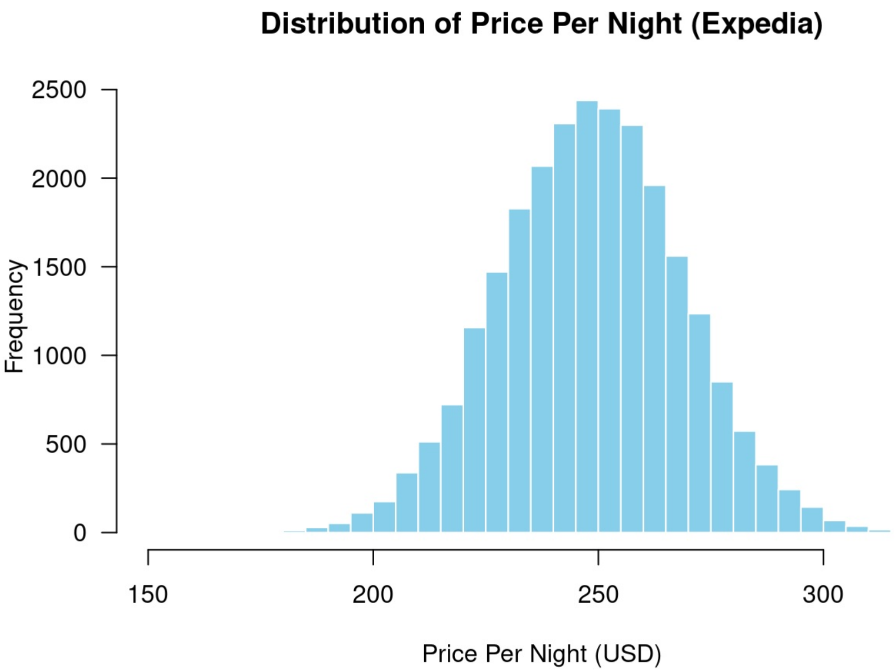
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
load("HW1.Rdata")

# Histogram:
hist(
  Expedia$PricePerNight,
  breaks = 50,
  col = "skyblue",
  border = "white",
  main = "Distribution of Price Per Night (Expedia)",  # Descriptive title
  xlab = "Price Per Night (USD)",       # X-axis label
  ylab = "Frequency",                   # Y-axis label
  xlim = c(150, max(Expedia$PricePerNight, na.rm = TRUE)),  # Define x-axis range
  las = 1,                              # Rotate y-axis labels for readability
  freq = TRUE                           # Show frequency (not density)
)
```



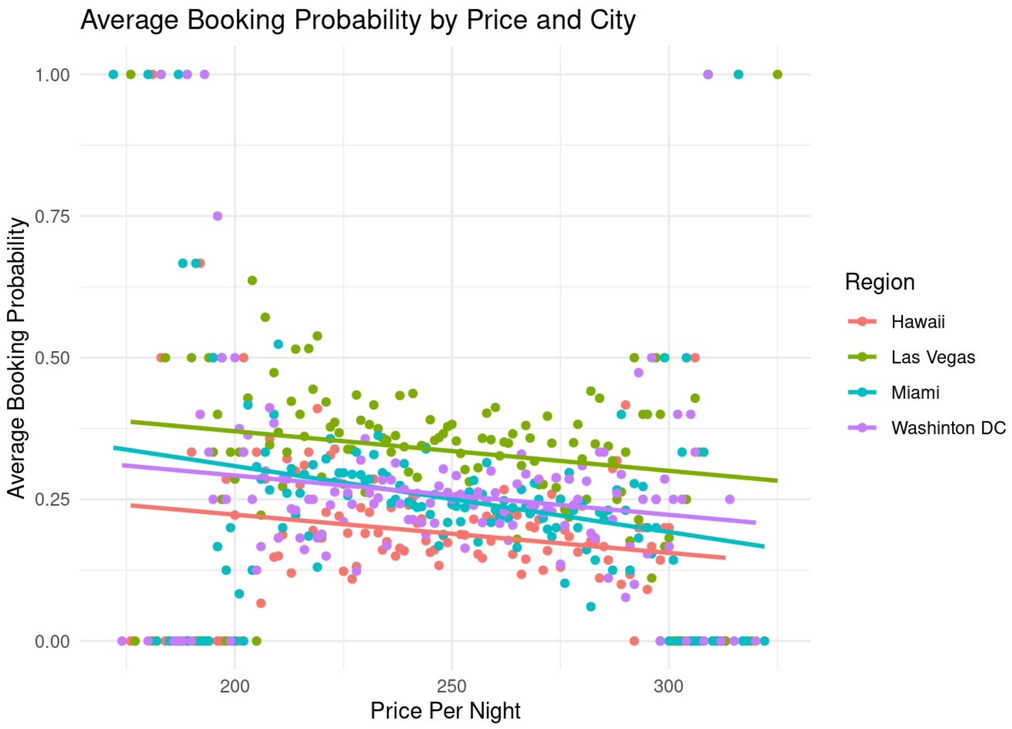
Model Free Analysis - Booked vs Price per night by Region

```
data_grouped <- Expedia %>%
  group_by(Region, PricePerNight) %>%
  summarize(AvgBooked = mean(`Booked?`))
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
ggplot(data_grouped, aes(x = PricePerNight, y = AvgBooked, color = Region)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Average Booking Probability by Price and City",
       x = "Price Per Night",
       y = "Average Booking Probability") +
  theme_minimal()
```

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



Model Free Analysis - Booked vs Price per night by User Income

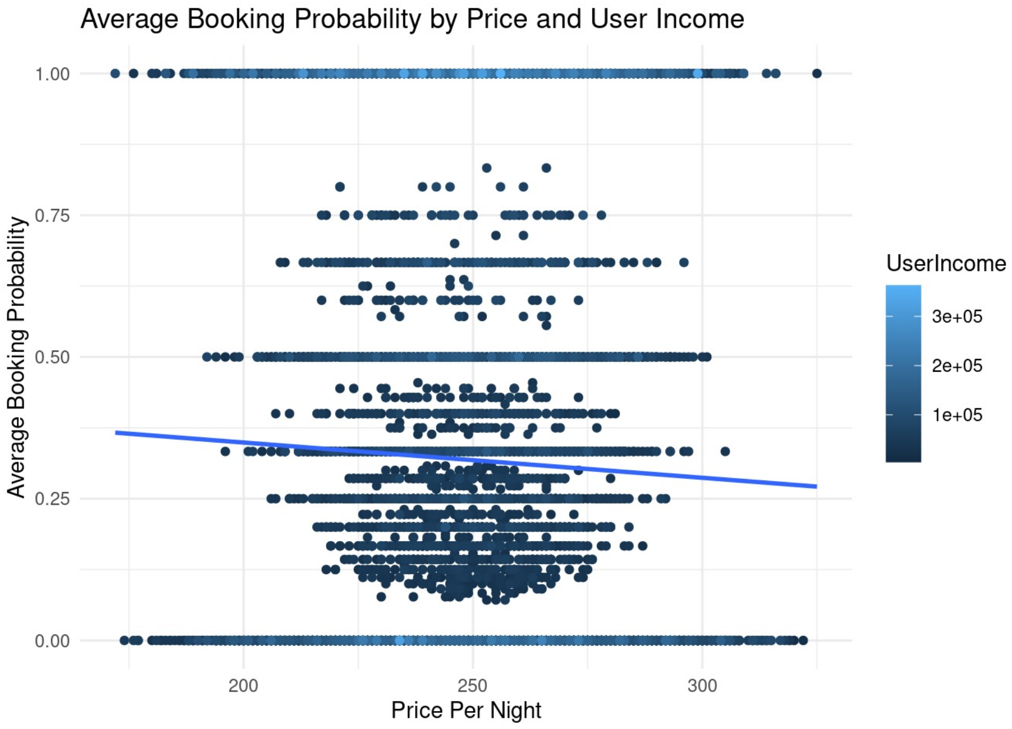
```
data_grouped <- Expedia %>%
  group_by(UserIncome, PricePerNight) %>%
  summarize(AvgBooked = mean(`Booked?`))
```

```
## `summarise()` has grouped output by 'UserIncome'. You can override using the
## `.groups` argument.
```

```
ggplot(data_grouped, aes(x = PricePerNight, y = AvgBooked, color = UserIncome)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Average Booking Probability by Price and User Income",
       x = "Price Per Night",
       y = "Average Booking Probability") +
  theme_minimal()
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Model - Simple linear Regression - Against Price Per Night

```
colnames(Expedia)
```

```
## [1] "PricePerNight" "Region"        "UserIncome"    "Booked?"
## [5] "Nights"
```

```
colnames(Expedia)[colnames(Expedia) == "Booked?"] <- "Booked"
## Plot 1: Simple Linear Regression Booked vs PricePerNight
model1 <- lm(Booked ~ PricePerNight, data = Expedia)
summary(model1)
```

```
##
## Call:
## lm(formula = Booked ~ PricePerNight, data = Expedia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3159 -0.2657 -0.2522  0.7126  0.7973
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.4463765  0.0342230  13.043  < 2e-16 ***
## PricePerNight -0.0007498  0.0001371  -5.471 4.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4383 on 24998 degrees of freedom
## Multiple R-squared:  0.001196,   Adjusted R-squared:  0.001156
## F-statistic: 29.93 on 1 and 24998 DF,  p-value: 4.52e-08
```

```
## Multiple R-squared:  0.001196,   Adjusted R-squared:  0.001156
## Plot 2: Simple Linear Regression Nights vs PricePerNight
model2 <- lm(Nights ~ PricePerNight, data = Expedia)
summary(model2)
```

```
##
## Call:
## lm(formula = Nights ~ PricePerNight, data = Expedia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9837 -0.7900 -0.7380  1.1522  5.2331
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.4867377  0.1054292  14.102  < 2e-16 ***
## PricePerNight -0.0028909  0.0004222  -6.847 7.72e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.35 on 24998 degrees of freedom
## Multiple R-squared:  0.001872,   Adjusted R-squared:  0.001832
## F-statistic: 46.88 on 1 and 24998 DF,  p-value: 7.718e-12
```

```
## Multiple R-squared:  0.001872,   Adjusted R-squared:  0.001832
```

Model - Multiple linear Regression - Booked vs PricePerNight

Dependent variable is Booked

```
## Add region and income
model_all <- lm(Booked ~ PricePerNight + UserIncome + as.factor(Region), data=Expedia)
summary(model_all)
```

```
##
## Call:
## lm(formula = Booked ~ PricePerNight + UserIncome + as.factor(Region),
##     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 ***
## UserIncome                     4.045e-06  8.512e-08  47.517  < 2e-16 ***
## as.factor(Region)Las Vegas     1.488e-01  7.453e-03  19.965  < 2e-16 ***
## as.factor(Region)Miami         4.635e-02  7.453e-03   6.219 5.10e-10 ***
## as.factor(Region)Washinton DC  4.613e-02  7.453e-03   6.189 6.15e-10 ***
## ---
## 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
```

```
## Multiple R-squared:  0.09758,    Adjusted R-squared:  0.0974
```

Model - Multiple linear Regression - Nights vs PricePerNight

Dependent variable is Nights

```
## Add region and income
model2_all <- lm(Nights ~ PricePerNight + UserIncome + as.factor(Region), data=Expedia)
summary(model2_all)
```

```
##
## Call:
## lm(formula = Nights ~ PricePerNight + UserIncome + as.factor(Region),
##     data = Expedia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0301 -0.7841 -0.5044  0.3240  4.9445
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    7.376e-01  1.024e-01   7.206 5.93e-13 ***
## PricePerNight                 -3.078e-03  4.033e-04  -7.633 2.38e-14 ***
## UserIncome                     1.183e-05  2.634e-07  44.900  < 2e-16 ***
## as.factor(Region)Las Vegas     4.503e-01  2.307e-02  19.520  < 2e-16 ***
## as.factor(Region)Miami         1.373e-01  2.307e-02   5.952 2.69e-09 ***
## as.factor(Region)Washinton DC  1.331e-01  2.307e-02   5.768 8.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.289 on 24994 degrees of freedom
## Multiple R-squared:  0.08975,    Adjusted R-squared:  0.08957
## F-statistic: 492.9 on 5 and 24994 DF,  p-value: < 2.2e-16
```

```
## Multiple R-squared:  0.08975,    Adjusted R-squared:  0.08957
```

Model - Multiple linear Regression - Booked vs PricePerNight

Dependent variable is Booked

```
## Add region and income
model_all <- lm(Booked ~ PricePerNight + UserIncome + as.factor(Region), data=Expedia)
summary(model_all)
```

```
##
## Call:
## lm(formula = Booked ~ PricePerNight + UserIncome + as.factor(Region),
##     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 ***
## UserIncome                     4.045e-06  8.512e-08  47.517  < 2e-16 ***
## as.factor(Region)Las Vegas     1.488e-01  7.453e-03  19.965  < 2e-16 ***
## as.factor(Region)Miami         4.635e-02  7.453e-03   6.219 5.10e-10 ***
## as.factor(Region)Washinton DC  4.613e-02  7.453e-03   6.189 6.15e-10 ***
## ---
## 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
```

```
## Multiple R-squared:  0.09758,    Adjusted R-squared:  0.0974
```

Model - Multiple linear Regression - Nights vs PricePerNight

Dependent variable is Nights

```
## Add region and income
model2_all <- lm(Nights ~ PricePerNight + UserIncome + as.factor(Region), data=Expedia)
summary(model2_all)
```

```
##
## Call:
## lm(formula = Nights ~ PricePerNight + UserIncome + as.factor(Region),
##     data = Expedia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0301 -0.7841 -0.5044  0.3240  4.9445
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    7.376e-01  1.024e-01   7.206 5.93e-13 ***
## PricePerNight                 -3.078e-03  4.033e-04  -7.633 2.38e-14 ***
## UserIncome                     1.183e-05  2.634e-07  44.900  < 2e-16 ***
## as.factor(Region)Las Vegas     4.503e-01  2.307e-02  19.520  < 2e-16 ***
## as.factor(Region)Miami         1.373e-01  2.307e-02   5.952 2.69e-09 ***
## as.factor(Region)Washinton DC  1.331e-01  2.307e-02   5.768 8.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.289 on 24994 degrees of freedom
## Multiple R-squared:  0.08975,    Adjusted R-squared:  0.08957
## F-statistic: 492.9 on 5 and 24994 DF,  p-value: < 2.2e-16
```

```
## Multiple R-squared:  0.08975,    Adjusted R-squared:  0.08957
```