HW1 partb Q1

5/19/2020

The dataset was collected from Airbnb with data on listings in the city of Asheville, NC. Here is the data provided for each listing: • room id: A unique number identifying an Airbnb listing. • host id: A unique number identifying an Airbnb host. • room type: One of ‘Entire home/apt’, ‘Private room’, or ‘Shared room’ • reviews: The number of reviews that a listing has received. • overall satisfaction: The average rating (out of five) that the listing has received from those visitors who left a review. • accommodates: The number of guests a listing can accommodate. • bedrooms: The number of bedrooms a listing offers. • price: The price (in USD) for a night stay.

```

Loading the data

```
# Load data and print head
airbnb_data <- read.csv("airbnb_data.csv",header = TRUE)
head(airbnb_data)
```

```
##    room_id survey_id   host_id   room_type      city reviews
## 1 15771735      1498 101992409 Shared room Asheville       0
## 2 18284194      1498 126414164 Shared room Asheville      32
## 3 18091012      1498 122380971 Shared room Asheville       4
## 4 12286328      1498    746673 Shared room Asheville      24
## 5   156926      1498    746673 Shared room Asheville     152
## 6 12989718      1498    746673 Shared room Asheville      20
##   overall_satisfaction accommodates bedrooms price
## 1                  0.0            4        1    67
## 2                  5.0            4        1    76
## 3                  4.5            2        1    45
## 4                  4.5            6        1    26
## 5                  4.5            6        1    26
## 6                  4.5            4        1    26
```

Question 1: Fit a multiple linear regression model using price as the response variable and all others as predictor variables (Note: remove ‘id’ columns). Which variables are statistically significant in determining the price?

```
# create model using 'lm' and print summary
model1 <- lm(price ~ room_type + reviews + overall_satisfaction + accommodates + bedrooms, data
 = airbnb_data)
summary(model1)
```
```

```
##
## Call:
## lm(formula = price ~ room_type + reviews + overall_satisfaction +
##     accommodates + bedrooms, data = airbnb_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -367.8  -49.2    3.2   38.6 4032.7
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -23.36172   21.88618  -1.067  0.28609
## room_typePrivate room  -0.93115   13.21827  -0.070  0.94386
## room_typeShared room  -76.66780   59.90939  -1.280  0.20099
## reviews                 0.01090    0.09982   0.109  0.91310
## overall_satisfaction  -10.48160    3.47320  -3.018  0.00262 **
## accommodates           23.00721    5.23952   4.391 1.27e-05 ***
## bedrooms               85.64533   11.45983   7.474 1.95e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 167.1 on 847 degrees of freedom
## Multiple R-squared:  0.3228, Adjusted R-squared:  0.318
## F-statistic:  67.3 on 6 and 847 DF,  p-value: < 2.2e-16
```

Statistically significant variables: overall_satisfaction, accommodates, bedrooms (can be concluded from p-values in the summary)

Question 2: Interpret the coefficients for predictors: room type(Shared Room), bedrooms?

Interpretations are as follows: 1) Room type(Shared Room): Holding all other variables constant, a listing for a shared room has an estimated price of 76.67 USD less than an entire home/apt. 2) Bedrooms: Holding all other variables constant, the estimated price of a listing increases by 85.64 USD with an inremental bedroom in the property.

Question 3: Predict the price (nearest dollar) for a listing with the following factors: bedrooms = 1, accommodates = 2, reviews = 70, overall_satisfaction = 4, and room_type= ‘Private room’.

```
# prepare prediction data and use 'predict' to find the value
pred_data = data.frame(bedrooms = 1, accommodates = 2, reviews = 70, overall_satisfaction = 4, r
oom_type = 'Private room')
predict(model1, pred_data)
```

```
##        1
## 66.20316
```

The estimated price for such a listing is 66 dollars.

Question 4: Identify outliers using Cook’s distance approach. Remove points having Cook’s distance > 1. Rerun the model after removal of these points and print summary.

```
# Use cook's distance to identify outliers
cooks <-cooks.distance(model1)
which(cooks>1)
```

```
## 94 95
## 94 95
```

```
# remove the outliers
airbnb_data_2 = airbnb_data[-c(94,95),]

# creating new model and print summary
model2 <- lm(price ~ room_type + reviews + overall_satisfaction + accommodates + bedrooms, data
 = airbnb_data_2)
summary(model2)
```

```
##
## Call:
## lm(formula = price ~ room_type + reviews + overall_satisfaction +
##     accommodates + bedrooms, data = airbnb_data_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190.95  -32.43   -7.09   20.35  876.26
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             75.01310    9.09152   8.251 6.01e-16 ***
## room_typePrivate room -32.28201    5.38034  -6.000 2.92e-09 ***
## room_typeShared room  -91.69951   24.28958  -3.775 0.000171 ***
## reviews                -0.05915    0.04047  -1.462 0.144202
## overall_satisfaction   -6.78957    1.41118  -4.811 1.78e-06 ***
## accommodates           11.90698    2.14267   5.557 3.68e-08 ***
## bedrooms               35.93177    4.87968   7.364 4.25e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.73 on 845 degrees of freedom
## Multiple R-squared:  0.4249, Adjusted R-squared:  0.4208
## F-statistic:   104 on 6 and 845 DF,  p-value: < 2.2e-16
```