บทนำ

ข้อมูลใน final.csv ประกอบไปด้วย 18 Attributes และข้อมูล 199 ตัวอย่าง เกี่ยวกับข้อมูลสมมติของลูกค้าสถาบันทางการเงิน โดยรายละเอียดแต่ละ attribute แสดงดังตารางที่ 1

ตารางที่ 1 แสดงรายละเอียดข้อมูลแต่ละ attribute

ชื่อตัวแปร	คำอธิบาย
Attrition_Flag	สถานะของลูกค้า
Customer_Age	อายุ
Gender	เพศ
Dependent_count	จำนวนผู้อยู่ในความอุปถัมภ์
Education_Level	ระดับการศึกษา
Marital_Status	สถานภาพ
Income_Category	ระดับรายได้
Card_Category	ประเภทของบัตรเครดิตที่ถือ
Months_on_book	จำนวนเดือนที่ใช้บริการ
Total_Relationship_Count	จำนวนของประเภทผลิตภัณฑ์/บริการที่มีสถาบันการเงิน
Credit_Limit	วงเงินรวม
Total_Revolving_Bal	วงเงินสินเชื่อหมุนเวียน
Total_Trans_Amt	มูลค่ารวมของทุกรายการที่ใช้จ่าย
Total_Trans_Ct	จำนวนธุรกรรมที่ใช้จ่าย
Avg_Utilization_Ratio	อัตราการใช้วงเงิน

Part 1: Chi-Square Test of Independence

บทนำ

ข้อมูลที่นำมาใช้ในการทดสอบ Chi-Square Test of Independence ประเภท categorical ในชุดข้อมูลนี้ ประกอบไปด้วย 6 จาก 18 Attributes แสดงดัง ตารางที่ 2 โดยจะตั้งเป้าหมายที่ [Income_Categories] กับ attributes อื่นที่เหลือเพื่อทดสอบ Independence ของแต่ละข้อมูลเมื่อเทียบกับข้อมูล [Income_Categories] เนื่องจากต้องการพิจารณาข้อมูลแต่ละชุดว่าชุดใดที่ไม่เป็นอิสระต่อเป้าหมาย หรือมีความสัมพันธ์อย่างมีนัยสำคัญต่อข้อมูลเป้าหมาย

ตารางที่ 2 แสดงข้อมูล categorical ในชุดข้อมูล

Attributes	**Data**
Attrition_Flag	"Attrited Customer" "Existing Customer"
Gender	"M" "F"
Education_Level	"Uneducated" "College" "Graduate" "High School" "Post-Graduate" "Doctorate" "Unknown"
Marital_Status	"Single" "Married" "Divorced" "Unknown"
Income_Category	"Gold" "Blue" "Silver"
Card_Category	"$80K - $120K" "$40K - $60K" "Less than $40K" "$60K - $80K" "$120K +" "Unknown"

ผลการวิเคราะห์

สมมติฐานโดยจะทำการทำสอบ Chi-Square Test of Independence กำหนดเป้าหมาย [Income_Categories] กับ attributes อื่นที่เหลือ คำสั่งที่ใช้ในโปรแกรมสถิติ (R) ดังรูปที่ 1 โดยสมมติฐานมีดังต่อไปนี้

1. H_o: [Income_Categories] กับ [Attrition_Flag] are independent

   H_a: [Income_Categories] กับ [Attrition_Flag] are not independent

2. H_o: [Income_Categories] กับ [Gender] are independent

   H_a: [Income_Categories] กับ [Gender] are not independent

3. H_o: [Income_Categories] กับ [Education_Level] are independent

   H_a: [Income_Categories] กับ [Education_Level] are not independent

4. H_o: [Income_Categories] กับ [Marital_Status] are independent

   H_a: [Income_Categories] กับ [Marital_Status] are not independent

5. H_o: [Income_Categories] กับ [Card_Category] are independent

   H_a: [Income_Categories] กับ [Card_Category] are not independent

```
library(readr)
path <- "C://Final.csv"
df <- read_csv(path)

# cross tap
cross_tap_att_inc <- table(df$Attrition_Flag, df$Income_Category)
cross_tap_gen_inc <- table(df$Gender, df$Income_Category)
cross_tap_edu_inc <- table(df$Education_Level, df$Income_Category)
cross_tap_mar_inc <- table(df$Marital_Status, df$Income_Category)
cross_tap_card_inc <- table(df$Card_Category, df$Income_Category)

# Chi-sqr test
chisq.test(cross_tap_att_inc)
chisq.test(cross_tap_gen_inc)
chisq.test(cross_tap_edu_inc)
chisq.test(cross_tap_mar_inc)
chisq.test(cross_tap_card_inc)
```

รูปที่ 1 คำสั่งที่ใช้ในโปรแกรมสถิติ (R) สำหรับ Chi-Square Test of Independence

ผลของคำสั่งแสดง cross-tab ของแต่ละ attribute คู่กับ [Income_Categories] แสดงดัง รูปที่ 2 ถึง รูปที่ 6

```
>table(df$Attrition_Flag, df$Income_Category)

                    $120K + $40K - $60K $60K - $80K $80K - $120K Less than $40K Unknown
  Attrited Customer       1          4          1           4             26       2
  Existing Customer       1         26         15          19             83      17
```

รูปที่ 2 ผลของ cross-tab : [Income_Categories] กับ [Attrition_Flag]

```
>table(df$Gender, df$Income_Category)

    $120K + $40K - $60K $60K - $80K $80K - $120K Less than $40K Unknown
  F       0          25          0           0            104      19
  M       2           5         16          23              5       0
```

รูปที่ 3 ผลของ cross-tab : [Income_Categories] กับ [Gender]

```
>table(df$Education_Level, df$Income_Category)

                $120K + $40K - $60K $60K - $80K $80K - $120K Less than $40K Unknown
  College             0          4          3           2             14       1
  Doctorate           0          5          0           2              7       1
  Graduate            0          5          4           6             30       7
  High School         1          5          2           5             23       4
  Post-Graduate       0          5          2           0              8       1
  Uneducated          0          2          4           3              8       5
  Unknown             1          4          1           5             19       0
```

รูปที่ 4 ผลของ cross-tab : [Income_Categories] กับ [Education_Level]

```
>table(df$Marital_Status, df$Income_Category)

           $120K + $40K - $60K $60K - $80K $80K - $120K Less than $40K Unknown
  Divorced       0           1           2            2              7       5
  Married        2          16           4           11             41       8
  Single         0           9           8           10             50       5
  Unknown        0           4           2            0             11       1
```

รูปที่ 5 ผลของ cross-tab : [Income_Categories] กับ [Marital_Status]

```
>table(df$Card_Category, df$Income_Category)

         $120K + $40K - $60K $60K - $80K $80K - $120K Less than $40K Unknown
  Blue         2          29          16           22             108      19
  Gold         0           0           0            1               0       0
  Silver       0           1           0            0               1       0
```

รูปที่ 6 ผลของ cross-tab : [Income_Categories] กับ [Card_Category]

ผลของคำสั่งแสดง Chi-Square Test ของแต่ละ attribute คู่กับ [Income_Categories] แสดงดัง

```
> chisq.test(cross_tap_att_inc)

        Pearson's Chi-squared test

data:  cross_tap_att_inc
X-squared = 6.1335, df = 5, p-value = 0.2934
```

รูปที่ 7

```
> chisq.test(cross_tap_card_inc)

        Pearson's Chi-squared test

data:  cross_tap_card_inc
X-squared = 9.9231, df = 10, p-value = 0.4473
```

รูปที่ 11 จาก threshold α = 0.05 ตามสมมติฐานเป็นดังต่อไปนี้

1. [Income_Categories] กับ [Attrition_Flag] ; P-value = 0.29 : Fail to Reject H_o
2. [Income_Categories] กับ [Gender] ; P-value < 2.2e-16 : **Reject H_o**
3. [Income_Categories] กับ [Education_Level] ; P-value = 0.37 : Fail to Reject H_o
4. [Income_Categories] กับ [Marital_Status] ; P-value = 0.15 : Fail to Reject H_o
5. [Income_Categories] กับ [Card_Category] ; P-value = 0.44 : Fail to Reject H_o

```
> chisq.test(cross_tap_att_inc)

        Pearson's Chi-squared test

data:  cross_tap_att_inc
X-squared = 6.1335, df = 5, p-value = 0.2934
```

รูปที่ 7 ผลของ Chi-Square Test : [Income_Categories] กับ [Attrition_Flag]

```
> chisq.test(cross_tap_gen_inc)

        Pearson's Chi-squared test

data:  cross_tap_gen_inc
X-squared = 152.11, df = 5, p-value < 2.2e-16
```

รูปที่ 8 ผลของ Chi-Square Test : [Income_Categories] กับ [Gender]

```
> chisq.test(cross_tap_edu_inc)

        Pearson's Chi-squared test

data:  cross_tap_edu_inc
X-squared = 31.876, df = 30, p-value = 0.3733
```

รูปที่ 9 ผลของ Chi-Square Test : [Income_Categories] กับ [Education_Level]

```
> chisq.test(cross_tap_mar_inc)

        Pearson's Chi-squared test

data:  cross_tap_mar_inc
X-squared = 20.555, df = 15, p-value = 0.1516
```

รูปที่ 10 ผลของ Chi-Square Test : [Income_Categories] กับ [Marital_Status]

```
> chisq.test(cross_tap_card_inc)

        Pearson's Chi-squared test

data:  cross_tap_card_inc
X-squared = 9.9231, df = 10, p-value = 0.4473
```

รูปที่ 11 ผลของ Chi-Square Test : : [Income_Categories] กับ [Card_Category]

บทสรุป

ผลของคำสั่งแสดง Chi-Square Test ของแต่ละ attribute คู่กับ [Income_Categories] ที่ threshold α = 0.05 ตามสมมติฐานพบว่าเพียงแค่ผลของ Chi-Square Test : [Income_Categories] กับ [Gender] เท่านั้นที่สามารถปฏิเสธ H_o ด้วย P-value < 2.2e-16 หมายความว่า [Income_Categories] กับ [Gender] ไม่เป็นอิสระต่อกันอย่างมีนัยสำคัญทางสถิติ หรืออีกอย่างคือ [Income_Categories] กับ [Gender] ตัวแปรมีความสัมพันธ์หรือมีการเชื่อมโยงกัน

Part 2: Logit Model

บทนำ

ในการวิเคราะห์ด้วยแบบจำลอง Logit เพื่อทำนายสถานะของลูกค้า [Attrition_Flag] ที่มีตัวแปร "Attrited Customer" ลูกค้าที่หยุดหรือเลิกใช้บริการ และ "Existing Customer" ลูกค้าที่ยังใช้บริการ โดยสร้าง 2 แบบจำลองที่มีตัวแปรแตกต่างกัน โดยแบบจำลองที่ 1 จะใช้ตัวแปรเพื่อทำนายเป้าหมายจาก [Months_on_book] จำนวนเดือนที่ใช้บริการ, [Total_Relationship_Count] จำนวนผลิตภัณฑ์ และ[Total_Trans_Ct] จำนวนธุรกรรมที่ใช้จ่าย และแบบจำลองที่ 2 จะใช้ตัวแปร [Credit Limit] วงเงินรวม, [Total_Revolving_Bal] วงเงินสินเชื่อหมุนเวียน และ[Avg_Utilization_Ratio] อัตราการใช้วงเงินในการวิเคราห์ โดยตัวแปรทั้ง 2 แบบจำของแสดงดัง ตารางที่ 3

ตารางที่ 3 แสดงข้อมูลตัวแปรที่ใช้ในโมเดล

Model No.1		**Model No.2**	
Variables	**Target**	**Variables**	**Target**
Months_on_book	Attrition_Flag	Credit Limit	Attrition_Flag
Total_Relationship_Count		Total_Revolving_Bal	
Total_Trans_Ct		Avg_Utilization_Ratio	

ผลการวิเคราะห์

แบบจำลองที่ 1

สมมติฐานโดยจะทำการทำสอบ Logit Model เพื่อทำนายเป้าหมาย [Attrition_Flag] กับ attributes อื่น สำหรับแบบจำลองที่ 1 คำสั่งที่ใช้ในโปรแกรมสถิติ (R) รูปที่ 12 โดยมี cut-off ที่ 0.5 โดยสมมติฐานมีดังต่อไปนี้

1. H_o: $\beta_1 = 0$; [Attrition_Flag] กับ [Months_on_book] are no relationship
   H_a: $\beta_1 \neq 0$; [Attrition_Flag] กับ [Months_on_book] are relationship
2. H_o: $\beta_2 = 0$; [Attrition_Flag] กับ [Total_Relationship_Count] are no relationship
   H_a: $\beta_2 \neq 0$; [Attrition_Flag] กับ [Total_Relationship_Count] are relationship
3. H_o: $\beta_3 = 0$; [Attrition_Flag] กับ [Total_Trans_Ct] are no relationship
   H_a: $\beta_3 \neq 0$; [Attrition_Flag] กับ [Total_Trans_Ct] are relationship

```
library(readr)
path <- "C://Final.csv"
df <- read_csv(path)

# Assign parameters
df$att_tran<- ifelse(df$Attrition_Flag == "Attrited Customer",0, 1) #transform char to 0 1
target_att <- df$att_tran
var1_mb <- df$Months_on_book
var2_re <- df$Total_Relationship_Count
var3_tran_ct <- df$Total_Trans_Ct

# Log Reg
reg_log <- glm(target_att ~ var1_mb + var2_re +var3_tran_ct, family = binomial)
summary(reg_log)

# Predict
y_pred  <- predict(reg_log, type = "response") # transform to logistic


# CM
table(y_prd > 0.5, target_att)
```

รูปที่ 12 คำสั่งที่ใช้ในโปรแกรมสถิติ (R) สำหรับ Logit Model - 1

ผลการทดสอบของแบบจำลองที่ 1 จาก threshold α = 0.05 เป็นดัง รูปที่ 13 และ confusion matrix จากการเปรียบเทียบการทำนาย และค่าจริงจาก n= 199 samples TN = 34 และ TP = 158 ดังนั้นค่า accuracy ที่ได้คือ 96.48% ดัง รูปที่ 14 และตามสมมติฐานที่ตั้งไว้พบว่า

1. [Attrition_Flag] กับ [Months_on_book]; P-value = 0.37 : Fail to Reject H_o
2. [Attrition_Flag] กับ [Total_Relationship_Count]; P-value = 0.11 : Fail to Reject H_o
3. [Attrition_Flag] กับ [Total_Trans_Ct]; P-value = 2.84e-06: **Reject H_o**

```
> summary(reg_log)

Call:
glm(formula = target_att ~ var1_mb + var2_re + var3_tran_ct, family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -23.17847    5.57181  -4.160 3.18e-05 ***
var1_mb        0.06131    0.06922   0.886    0.376
var2_re        0.51961    0.32617   1.593    0.111
var3_tran_ct   0.34559    0.07381   4.682 2.84e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.067  on 198  degrees of freedom
Residual deviance:  34.208  on 195  degrees of freedom
AIC: 42.208

Number of Fisher Scoring iterations: 8
```

รูปที่ 13 ผลสรุปภาพรวมแบบจำลองที่ 1

```
> table(y_pred > 0.5, target_att)
       target_att
          0   1
  FALSE  34   3
  TRUE    4 158
```

รูปที่ 14 confusion matrix ของแบบจำลองที่ 1

แบบจำลองที่ 2

สมมติฐานโดยจะทำการทำสอบ Logit Model เพื่อทำนายเป้าหมาย [Attrition_Flag] กับ attributes อื่น สำหรับแบบจำลองที่ 2 คำสั่งที่ใช้ในโปรแกรมสถิติ (R) รูปที่ 15 โดยมี cut-off ที่ 0.5 โดยสมมติฐานมีดังต่อไปนี้

1. H_o: $\beta_1 = 0$; [Attrition_Flag] กับ [Credit Limit] are no relationship
   H_a: $\beta_1 \neq 0$; [Attrition_Flag] กับ [Credit Limit] are relationship
2. H_o: $\beta_2 = 0$; [Attrition_Flag] กับ [Total_Revolving_Bal] are no relationship
   H_a: $\beta_2 \neq 0$; [Attrition_Flag] กับ [Total_Revolving_Bal] are relationship
3. H_o: $\beta_3 = 0$; [Attrition_Flag] กับ [Avg_Utilization_Ratio] are no relationship
   H_a: $\beta_3 \neq 0$; [Attrition_Flag] กับ [Avg_Utilization_Ratio] are relationship

```
library(readr)
path <- "C://Final.csv"
df <- read_csv(path)

# Assign parameters
df$att_tran<- ifelse(df$Attrition_Flag == "Attrited Customer",0, 1) #transform char to 0 1
target_att <- df$att_tran
var1_cl <- df$Credit_Limit
var2_rev <- df$Total_Revolving_Bal
var3_avg <- df$Avg_Utilization_Ratio

# Log Reg
reg_log <- glm(target_att ~ var1_cl + var2_rev + var3_avg, family = binomial)
summary(reg_log)

# Predict
y_pred  <- predict(reg_log, type = "response") # transform to logistic


# CM
table(y_pred > 0.5, target_att)
```

รูปที่ 15 คำสั่งที่ใช้ในโปรแกรมสถิติ (R) สำหรับ Logit Model - 2

ผลการทดสอบของแบบจำลองที่ 1 จาก threshold α = 0.05 เป็นดัง รูปที่ 16 และ confusion matrix จากการเปรียบเทียบการทำนาย และค่าจริงจาก n= 199 samples TP = 161 ดังนั้นค่า accuracy ที่ได้คือ 80.90% ดังรูปที่ 17 และตามสมมติฐานที่ตั้งไว้พบว่า

1. [Attrition_Flag] กับ [Credit Limit]; P-value = 0.68 : Fail to Reject H_o
2. [Attrition_Flag] กับ [Total_Revolving_Bal]; P-value = 0.02 : **Reject H_o**
3. [Attrition_Flag] กับ [Avg_Utilization_Ratio]; P-value = 0.99: Fail to Reject H_o

```
> summary(reg_log)

Call:
glm(formula = target_att ~ var1_cl + var2_rev + var3_avg, family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.787e-01  3.223e-01   0.865   0.3872
var1_cl     2.322e-05  5.768e-05   0.403   0.6873
var2_rev    1.667e-03  7.471e-04   2.231   0.0257 *
var3_avg    3.478e-03  1.847e+00   0.002   0.9985
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.07  on 198  degrees of freedom
Residual deviance: 158.80  on 195  degrees of freedom
AIC: 166.8

Number of Fisher Scoring iterations: 5
```

รูปที่ 16 ผลสรุปภาพรวมแบบจำลองที่ 2

```
> table(y_pred > 0.5, target_att)
      target_att
         0   1
  TRUE  38 161
```

รูปที่ 17 confusion matrix ของแบบจำลองที่ 2

บทสรุป

สรุปผลจากการวิเคราะห์แบบจำลอง Logit 2 แบบจำลองกับเป้าหมาย [Attrition_Flag] พบว่าตัวแปรที่มีความสัมพันธ์หรือความเชื่องโยงในแบบจำลองที่ 1 คือ [Total_Trans_Ct] และสำหรับแบบจำลองที่ 2 คือ [Total_Revolving_Bal] โดยค่า accuracy ของแบบจำลองที่ 1 มากกว่า แบบจำลองที่ 2 แสดงดัง ตารางที่ 4

ตารางที่ 4 แสดงตัวแปรที่ใช้และค่า accuracy ที่ได้จากการทำนาย

Type	Model No.1	Model No.2
	Variables	Variables
Var 1	Months_on_book	Credit Limit
Var 2	Total_Relationship_Count	**Total_Revolving_Bal***
Var 3	**Total_Trans_Ct***	Avg_Utilization_Ratio
accuracy	96.48	80.90

** Relationship with target*

Part 3: Regression Analysis

บทนำ

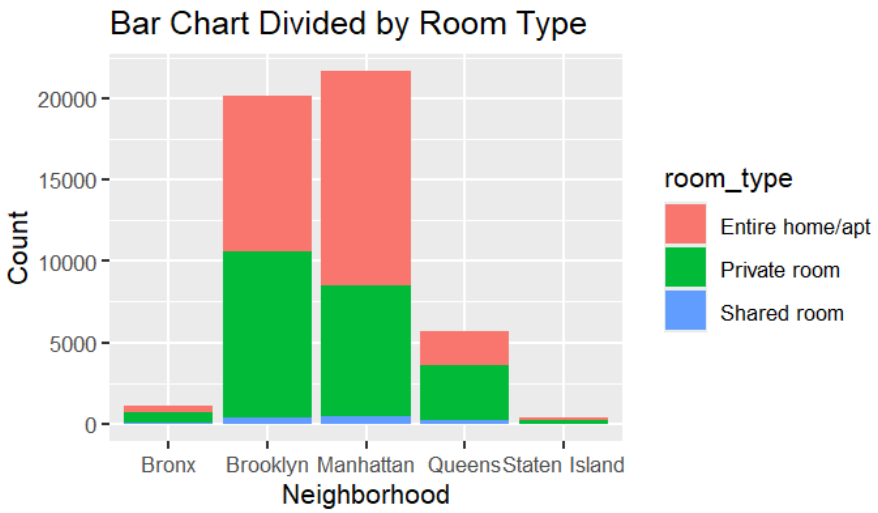
ข้อมูลเกี่ยวกับ New York City Airbnb ปี 2019 โดยเป็นข้อมูลการเปิดที่พักให้เช่าผ่าน Airbnb ในเมือง New York City ปี 2019 จาก Kaggle* ประกอบไปด้วยตัวแปรทั้ง 16 ตัว/คอลัมน์ ได้แก่

1. "id" : id ที่พัก
2. "name" : ชื่อที่พัก
3. "host_id" : id เจ้าของที่พัก
4. "host_name" : ชื่อเข้าของที่พัก
5. "neighbourhood_group" : ย่านของที่พัก/เมือง
6. "neighbourhood" : ชุมชนของย่านที่พัก/เมือง
7. "latitude" : พิกัด latitude
8. "longitude" : พิกัด longitude
9. "room_type" : ประเภทห้อง
10. "price": ราคาที่พัก
11. "minimum_nights": คืนขั้นต่ำในการเข้าพัก
12. "number_of_reviews" : จำนวนรีวิว
13. "last_review" : วันที่รีวิวล่าสุด
14. "reviews_per_month" : จำนวนรีวิวต่อเดือน
15. "calculated_host_listings_count" : จำนวนที่พักที่เจ้าของมีใน Airbnb
16. "availability_365" : เวลาที่เปิดให้บริการจาก 365 วัน

** https://www.kaggle.com/code/whyalwaysme/ab-nyc-2019/notebook?select=AB_NYC_2019.csv*

โดยข้อมูลที่จะใช้ในการวิเคราะห์จะมุ่งเน้นไปที่เมือง 'Manhattan' เนื่องจากเป็นเมืองที่มีข้อมูลมากที่สุด โดยสนใจประเภทของห้องพัก Private room และ Entire home/apt เนื่องจากปริมาณ Shared room มีน้อยมากเมื่อเทียบกับทั้งสองประเภท แสดงรายละเอียดดัง รูปที่ 18 โดยตัวแปรเป้าหมายในการวิเคราะห์คือ "price" และตัวแปรอิสระที่ใช้ในการวิเคราะห์ 5 ตัวแปรที่อาจจะส่งผลต่อราคาที่พักในย่าน 'Manhattan' โดยตัดข้อมูลทั่วไปเช่น id, ชื่อ, พิกัด รวมไปถึงวันที่รีวิวล่าสุด และจำนวนที่พักที่เจ้าของมีใน Airbnb โดย 5 ตัวแปรที่เลือกได้แก่

1. "room_type"
2. "minimum_nights"
3. "number_of_reviews"
4. "reviews_per_month"
5. "availability_365"



รูปที่ 18 chart แสดงประเภทที่พักตามเมืองใน New York

ผลการวิเคราะห์

สมมติฐานโดยจะทำการทำสอบ Regression Model เพื่อทำนายเป้าหมาย ["price"] กับ attributes ต่างๆ สำหรับคำสั่งที่ใช้ในโปรแกรมสถิติ (R) รูปที่ 19 โดยสมมติฐานมีดังต่อไปนี้

Overall Significant:

1. H_o: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$; all are no relationship

   H_a: At least one of slopes is non-zero ; at least 1 variable are relationship with target

Significance of Slope:

1. H_o: $\beta_1 = 0$; [“price”] กับ ["room_type"] are no relationship

   H_a: $\beta_1 \neq 0$; [“price”] กับ ["room_type"] are relationship
2. H_o: $\beta_2 = 0$; [“price”] กับ ["minimum_nights"] are no relationship

H_a: $\beta_2 \neq 0$; [“price”] กับ ["minimum_nights"] are relationship

3. H_o: $\beta_3 = 0$; [“price”] กับ ["number_of_reviews"] are no relationship

   H_a: $\beta_3 \neq 0$; [“price”] กับ ["number_of_reviews"] are relationship

4. H_o: $\beta_4 = 0$; [“price”] กับ ["reviews_per_month"] are no relationship

   H_a: $\beta_4 \neq 0$; [“price”] กับ ["reviews_per_month"] are relationship

5. H_o: $\beta_5 = 0$; [“price”] กับ ["availability_365"] are no relationship

   H_a: $\beta_5 \neq 0$; [“price”] กับ ["availability_365"] are relationship

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

path <- "C://AB_NYC_2019.csv"
df <- read_csv(path)

# Selected Manhattan and "Private room","Entire home/apt
# Eliminate null data
# Transform "Private room","Entire home/apt to 0, 1
df_selected <- filter(df, `neighbourhood_group` == 'Manhattan' &
                        `room_type` == c("Private room","Entire home/apt"))
df_selected <- na.omit(df_selected)
df_selected$room_type <- ifelse(df_selected$room_type == "Private room", 0, 1)

# Assing Variables
target_price <- df_selected$price
var1_rt <- df_selected$room_type
var2_mn <- df_selected$minimum_nights
var3_nr <- df_selected$number_of_reviews
var4_rm <- df_selected$reviews_per_month
var5_365 <- df_selected$availability_365

# Multiple Reg
reg_mul <- lm(target_price ~ var1_rt + var2_mn +var3_nr +var4_rm + var5_365)
summary(reg_mul)
```

รูปที่ 19 คำสั่งที่ใช้ในโปรแกรมสถิติ (R) สำหรับ Regression Analysis

ผลการวิเคราะห์สำหรับ Regression Analysis จาก threshold α = 0.05 แสดงดัง รูปที่ 20 และตามสมมติฐานที่ตั้งไว้พบว่า

Overall Significant:

1. p-value: < 2.2e-16 ; **Reject H_o** (At least one of slopes is non-zero)
2. Goodness-of-fit: Adjusted R^2 = 0.0945 (Model สามารถอธิบายได้ 9.45%)

Significance of Slope:

1. [“price”] กับ ["room_type"] ; p-value: < 2.2e-16 ; **Reject H_o**
2. [“price”] กับ ["minimum_nights"] p-value: 0.0358 ; **Reject H_o**
3. [“price”] กับ ["number_of_reviews"] p-value: 3.04e-06 ; **Reject H_o**
4. [“price”] กับ ["reviews_per_month"] p-value: 0.0206 ; **Reject H_o**
5. [“price”] กับ ["availability_365"] p-value: < 2e-16 ; **Reject H_o**

```
> summary(reg_mul)

Call:
lm(formula = target_price ~ var1_rt + var2_mn + var3_nr + var4_rm + var5_365)

Residuals:
   Min     1Q Median     3Q    Max
-224.7  -72.2  -24.3   18.6 9917.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.72083    4.65624  18.195  < 2e-16 ***
var1_rt     118.06074    4.85735  24.306  < 2e-16 ***
var2_mn      -0.24660    0.11744  -2.100   0.0358 *
var3_nr      -0.28193    0.06035  -4.672 3.04e-06 ***
var4_rm       4.13891    1.78770   2.315   0.0206 *
var5_365      0.27137    0.01914  14.175  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 210.8 on 8058 degrees of freedom
Multiple R-squared:  0.09506,Adjusted R-squared:  0.0945
F-statistic: 169.3 on 5 and 8058 DF,  p-value: < 2.2e-16
```

รูปที่ 20 ผลสรุปภาพรวมของ Regression Analysis

บทสรุป

สรุปผลจากการวิเคราะห์ Regression Analysis กับเป้าหมาย [“price”] พบว่าแบบจำลองมีอย่างน้อย 1 ตัวแปรที่ความสัมพันธ์ และค่า R^2 = 0.0945 หมายความว่าถึงแม้แบบจำลองกับตัวแปรจะมีความสัมพันธ์ แต่แบบจำลองยังไม่สามารถอธิบายได้อย่างน่าพอใจ เนื่องจากสามารถอธิบายตัวแปรในแบบจำลองได้ 9.45% และเมื่อตรวจสอบแต่ละตัวแปร พบว่าตัวแปรทุกตัวที่มีความสัมพันธ์หรือความเชื่องโยงในแบบจำลอง Regression