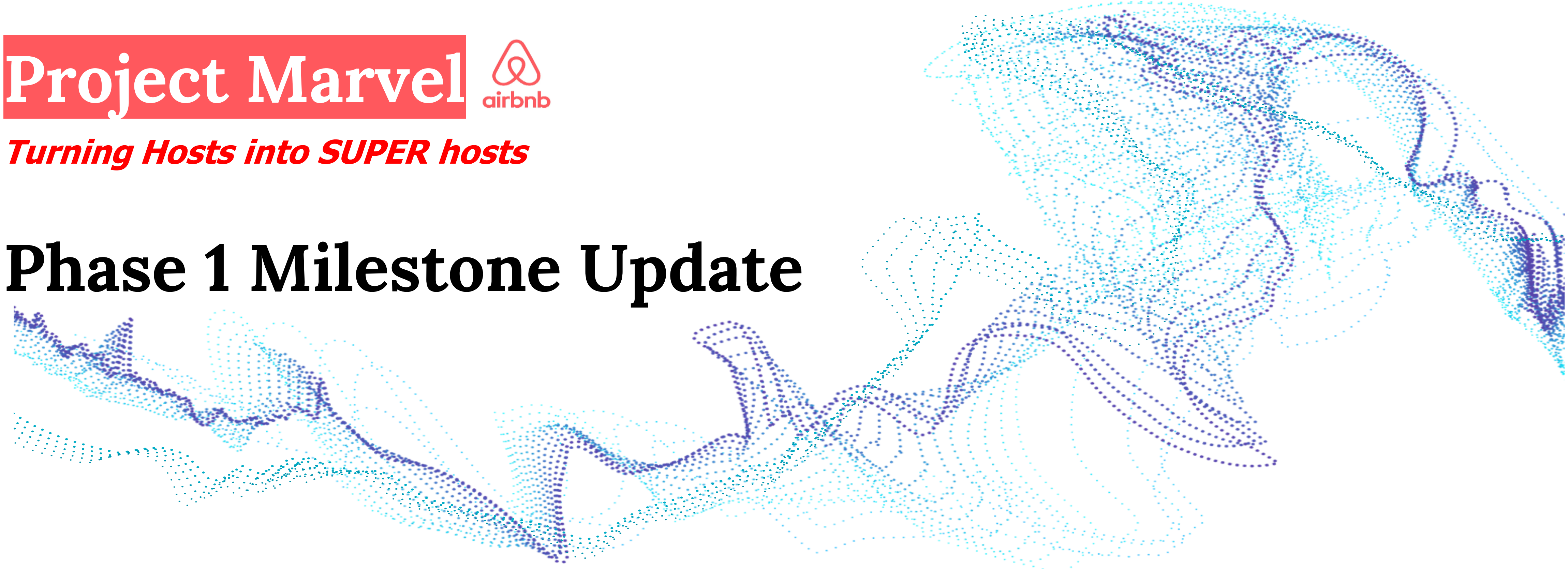
Group18 Consulting Co.

Project Marvel

airbnb

Turning Hosts into SUPER hosts

Phase 1 Milestone Update



G18 Consulting Project Team:

Jonathan YORK
Jisoo KIM
Daoqing SU
Anthony WONG
Elsa ZHAN
Terence ZHANG

G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

1.
Introduction

- Definition of the project
- Background of the Airbnb's 'Superhost' programme
- Hong Kong Market Study

1.1 Project Definition

What are some simple and actionable factors that can help hosts in Hong Kong become 'Superhosts'? 🎖️

Quantitative analyses on the influential factors that **help hosts in Hong Kong to become "Superhosts"**

G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

1.2 Background – “Superhost”

What is a ‘Superhost’?

Airbnb’s top performing hosts.

The host must own an account in good standing who has met the following criteria in the past 12 months:

- Completed **at least 10 trips** or 3 reservations that total at least 100 nights;
- Maintained a **90% + response rate;**
- Maintained **a less than 1% cancellation rate**, with exceptions made for those that fall under Airbnb’s extenuating circumstances policy;
- Maintained a **4.8 overall rating**.

Why ‘Superhost’?

More visibility from prospective guests, additional earning potentials, exclusive rewards and getting priority support from Airbnb

- 5% increase in **weekly views**
- 81% higher **occupancy rate**
- Earn **60% more daily revenue** than regular hosts on average
- **Cash rewards** from AirBnB for mentorship

What’s in it for Airbnb?

Given that Airbnb revenue comes from two major sources, it is also incentivised to encourage more hosts to become “Superhosts”:

- **Commission from hosts:** Everytime someone chooses a host’s property and makes payment, Airbnb takes 10% of the payment amount as commission.
- **Transaction fee from travelers:** When travelers make payments for stays, they are charged a 3% fee for the transaction. This amount adds to the Airbnb revenue.

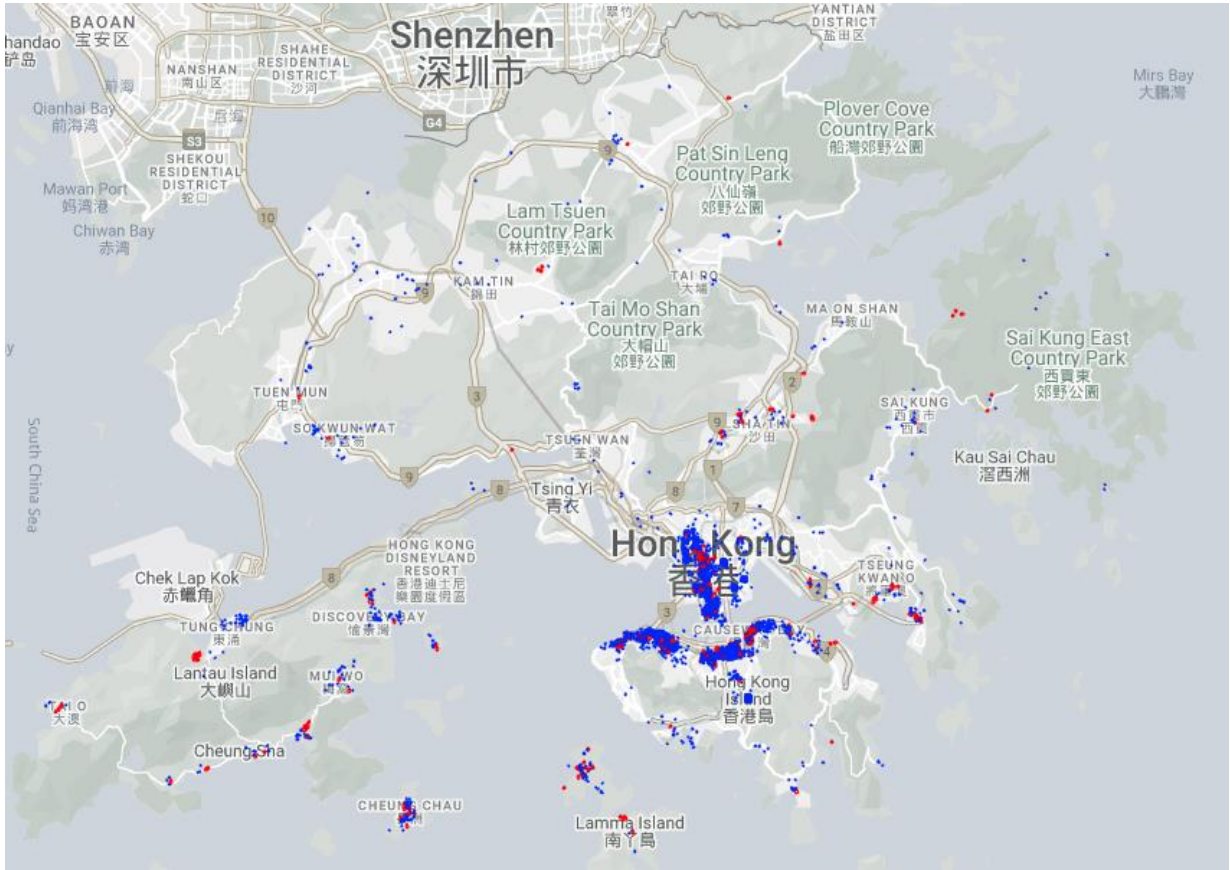
G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

1.2 Background - Market Study

🌐 19.4%

🇭🇰 10.5%

1.2 Background - Hong Kong Market Study



5056 listings, 532 superhosts

Data last scraped in 16 Sept 2022
Includes hosts from 2009 to 2022

Neighbourhood Impact?

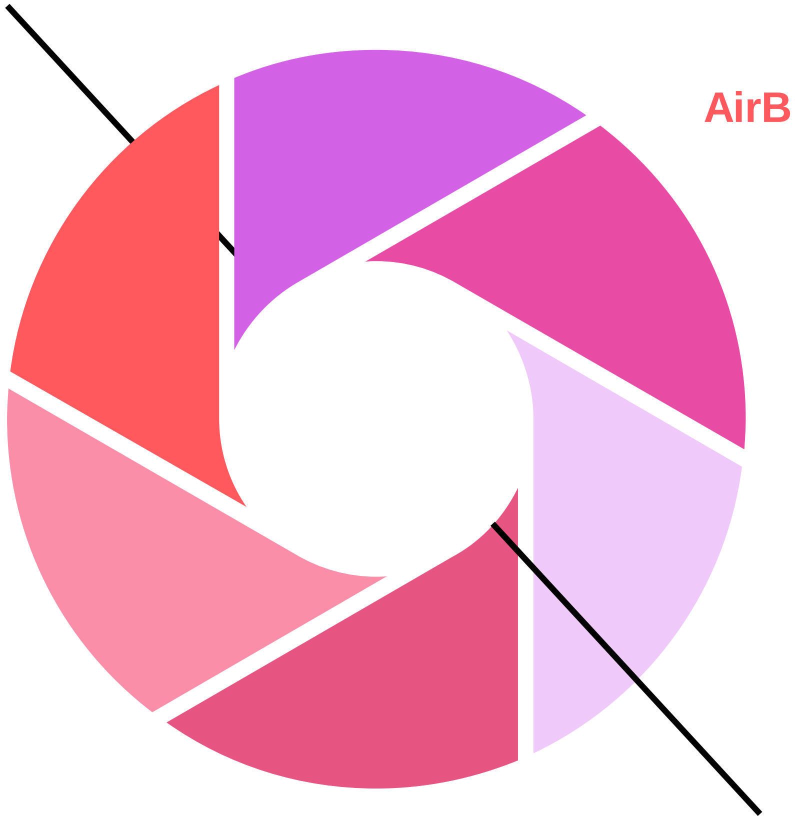
Listings **in Shatin, Sai Kung** and the **Hong Kong Island** are typically listed by “Superhosts”.

Listings in **Kwai Tsing, Wong Tai Sin** and **Sham Shui Po** have the lowest rate of “Superhosts”.

G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

1.3 Stakeholder Analysis

Providing practical tips to assist aspiring hosts and existing hosts becoming “Superhosts”.



Stakeholders to be influenced

Aspiring hosts

Existing & potential guests of Airbnb

AirBnb

Existing hosts

Stakeholders with influence

2.
Analysis & Findings

- Data Cleaning and Preparation
- Models and findings

2.1 Data Preparation

Dataset:

- 5056 listings in HK, 75 columns
- Overview on variables:

  - Information on host : location, id, verification status
  - Information on listing : max/min nights available, price, amenities, room type, location
  - Information on review scores : breakdown of review scores and its values

Steps taken for Data Preparation:

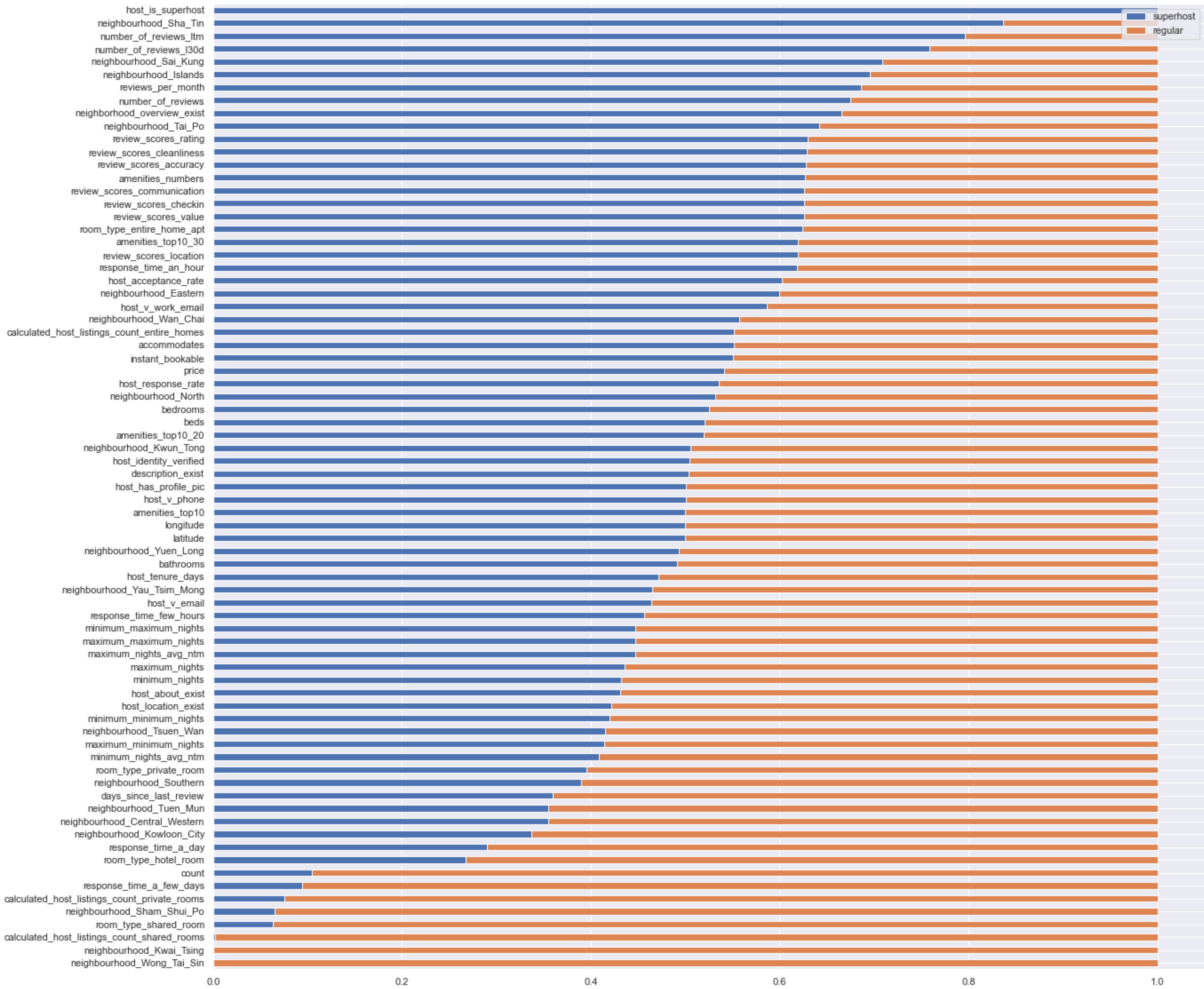
1. converting data with ‘object’ data type to appropriate data types;
2. dropping columns that have more than **75%** missing values;
3. filling missing data with appropriate entries;
4. dropping variables; and
5. creating categorical/dummy variables.

2.2 Understanding Host vs. Superhost

Understanding the differences

1. Neighbourhood matters!
2. Reviews and Superhosts are correlated
3. Number of reviews also matter to becoming Superhosts
4. …
5. There is a lot…

TLDR: There are many factors which correlate to becoming a Superhost, however, not all of them are actionable nor simple.

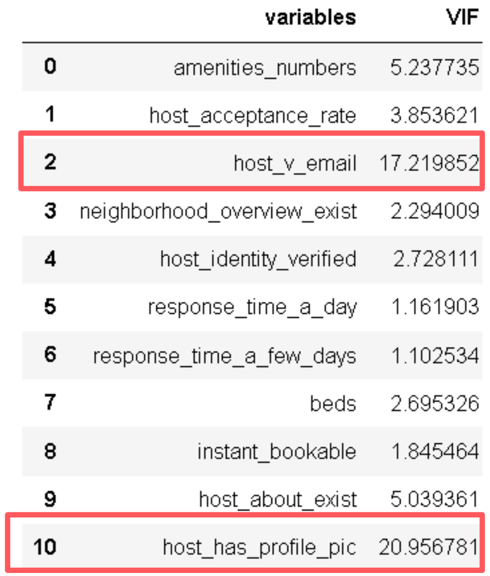


G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

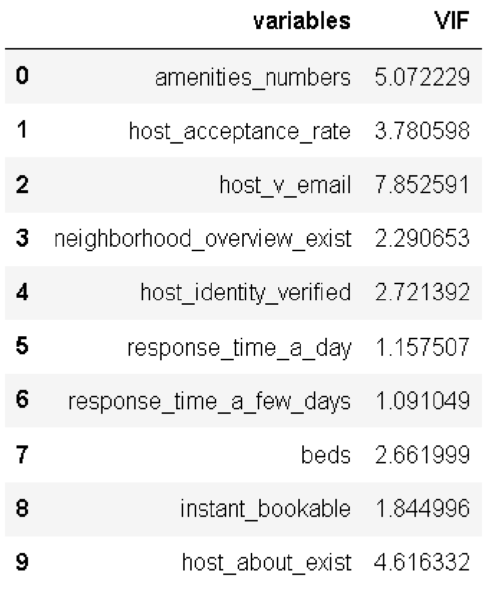
2.2 Models and Results

Model 1 : Logistic Regression on Feature Data

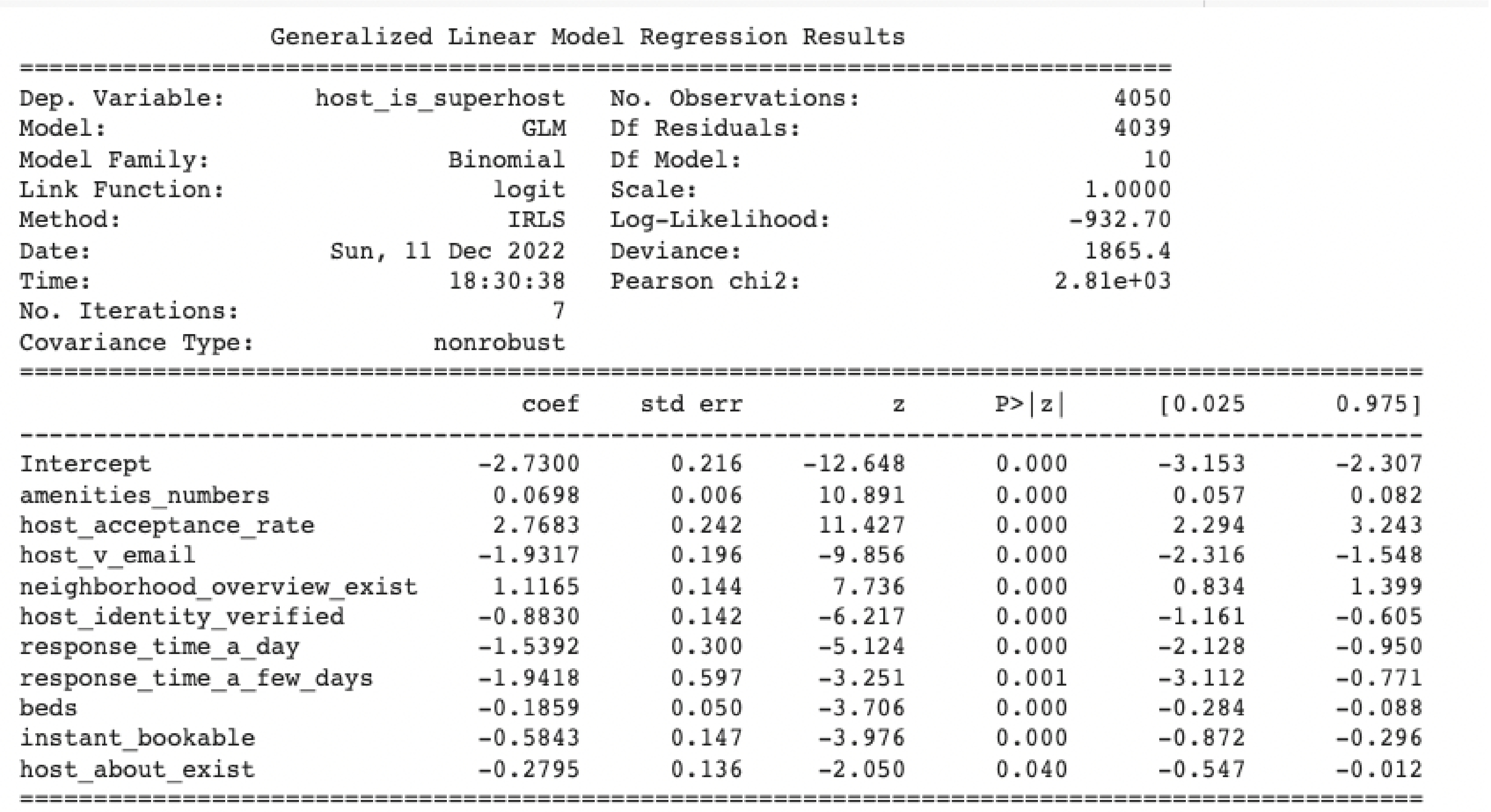
- Host_is_Superhost : Dependant Variable (Binary)
- Used AIC forward Selection model to further narrow down the number of variables.
- Conducted VIF analysis to check for multicollinearity issue. Where we decided to drop 'host_has_profilepic'
- Our final logistic regression model contains 9 variables.



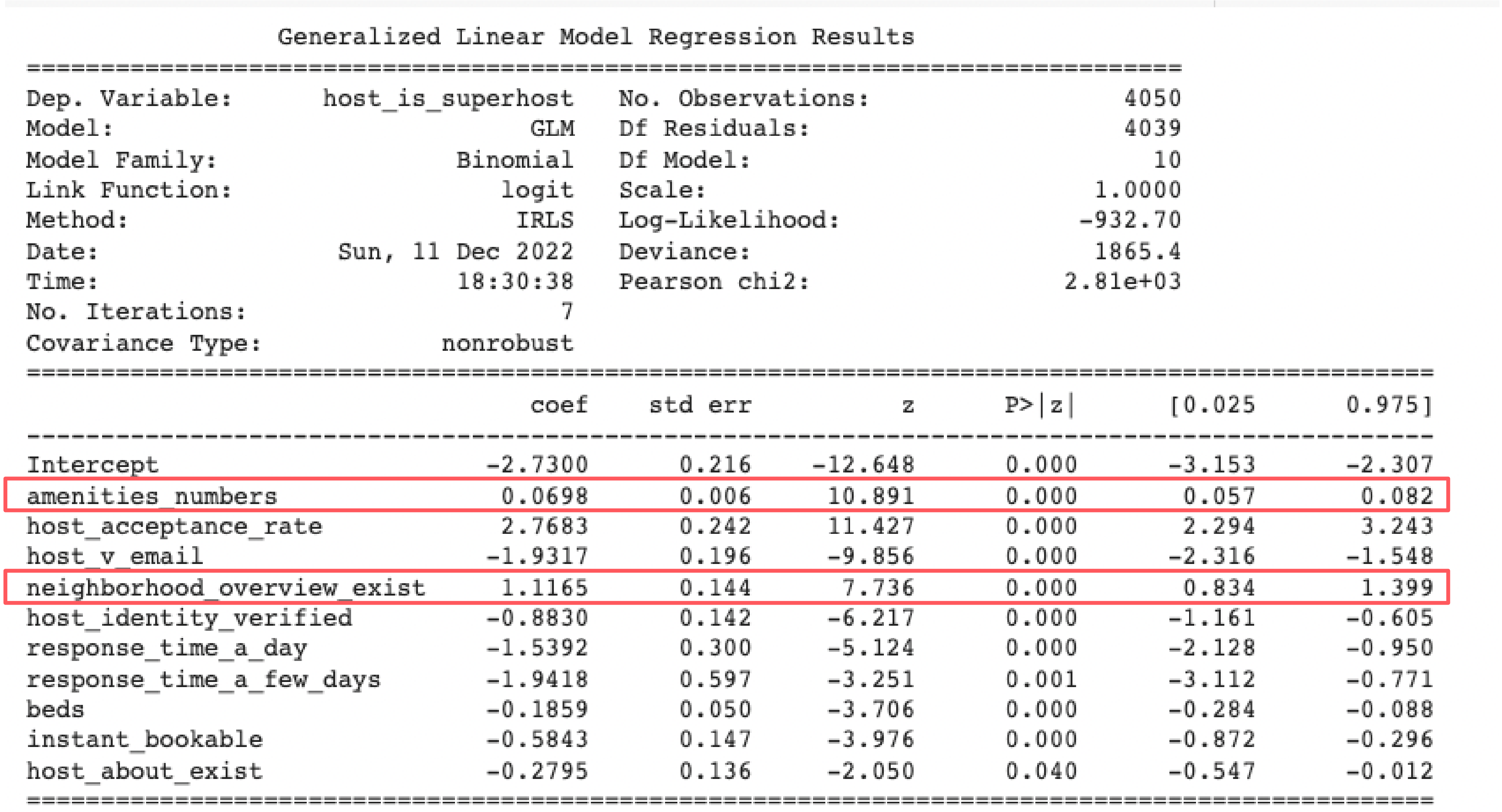
	variables	VIF
0	amenities_numbers	5.237735
1	host_acceptance_rate	3.853621
2	host_v_email	17.219852
3	neighborhood_overview_exist	2.294009
4	host_identity_verified	2.728111
5	response_time_a_day	1.161903
6	response_time_a_few_days	1.102534
7	beds	2.695326
8	instant_bookable	1.845464
9	host_about_exist	5.039361
10	host_has_profile_pic	20.956781



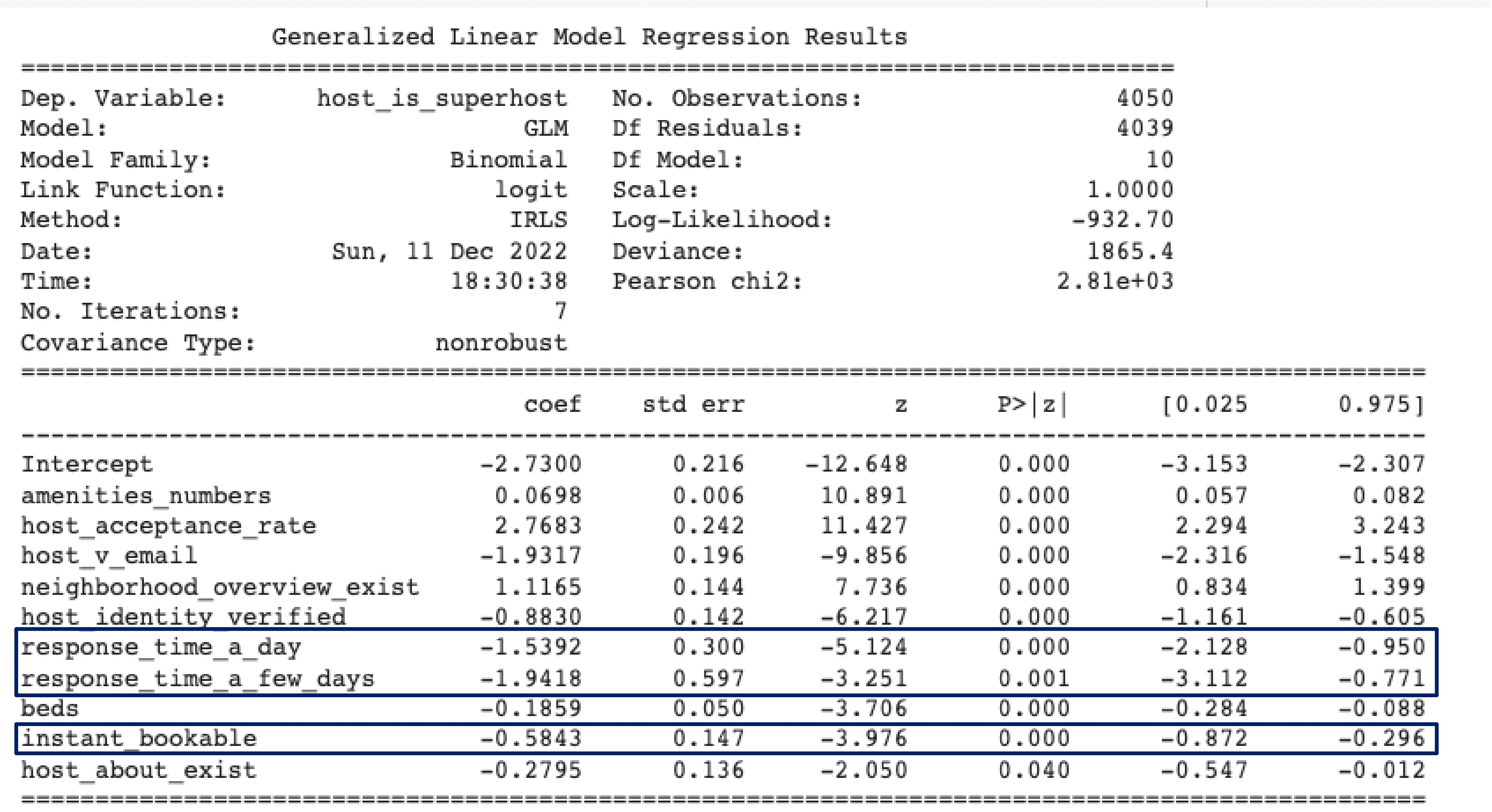
	variables	VIF
0	amenities_numbers	5.072229
1	host_acceptance_rate	3.780598
2	host_v_email	7.852591
3	neighborhood_overview_exist	2.290653
4	host_identity_verified	2.721392
5	response_time_a_day	1.157507
6	response_time_a_few_days	1.091049
7	beds	2.661999
8	instant_bookable	1.844996
9	host_about_exist	4.616332



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:      host_is_superhost   No. Observations:                 4050
Model:                            GLM   Df Residuals:                     4039
Model Family:                Binomial   Df Model:                           10
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -932.70
Date:                Sun, 11 Dec 2022   Deviance:                       1865.4
Time:                        18:30:38   Pearson chi2:                 2.81e+03
No. Iterations:                     7
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                      -2.7300      0.216    -12.648      0.000      -3.153      -2.307
amenities_numbers               0.0698      0.006     10.891      0.000       0.057       0.082
host_acceptance_rate            2.7683      0.242     11.427      0.000       2.294       3.243
host_v_email                   -1.9317      0.196     -9.856      0.000      -2.316      -1.548
neighborhood_overview_exist     1.1165      0.144      7.736      0.000       0.834       1.399
host_identity_verified         -0.8830      0.142     -6.217      0.000      -1.161      -0.605
response_time_a_day            -1.5392      0.300     -5.124      0.000      -2.128      -0.950
response_time_a_few_days       -1.9418      0.597     -3.251      0.001      -3.112      -0.771
beds                           -0.1859      0.050     -3.706      0.000      -0.284      -0.088
instant_bookable               -0.5843      0.147     -3.976      0.000      -0.872      -0.296
host_about_exist               -0.2795      0.136     -2.050      0.040      -0.547      -0.012
===============================================================================================
```



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:      host_is_superhost   No. Observations:                 4050
Model:                            GLM   Df Residuals:                     4039
Model Family:                Binomial   Df Model:                           10
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -932.70
Date:                Sun, 11 Dec 2022   Deviance:                       1865.4
Time:                        18:30:38   Pearson chi2:                 2.81e+03
No. Iterations:                     7
Covariance Type:            nonrobust
===============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                   -2.7300      0.216    -12.648      0.000      -3.153      -2.307
amenities_numbers            0.0698      0.006     10.891      0.000       0.057       0.082
host_acceptance_rate         2.7683      0.242     11.427      0.000       2.294       3.243
host_v_email                -1.9317      0.196     -9.856      0.000      -2.316      -1.548
neighborhood_overview_exist  1.1165      0.144      7.736      0.000       0.834       1.399
host_identity_verified      -0.8830      0.142     -6.217      0.000      -1.161      -0.605
response_time_a_day         -1.5392      0.300     -5.124      0.000      -2.128      -0.950
response_time_a_few_days    -1.9418      0.597     -3.251      0.001      -3.112      -0.771
beds                        -0.1859      0.050     -3.706      0.000      -0.284      -0.088
instant_bookable            -0.5843      0.147     -3.976      0.000      -0.872      -0.296
host_about_exist            -0.2795      0.136     -2.050      0.040      -0.547      -0.012
===============================================================================================
```



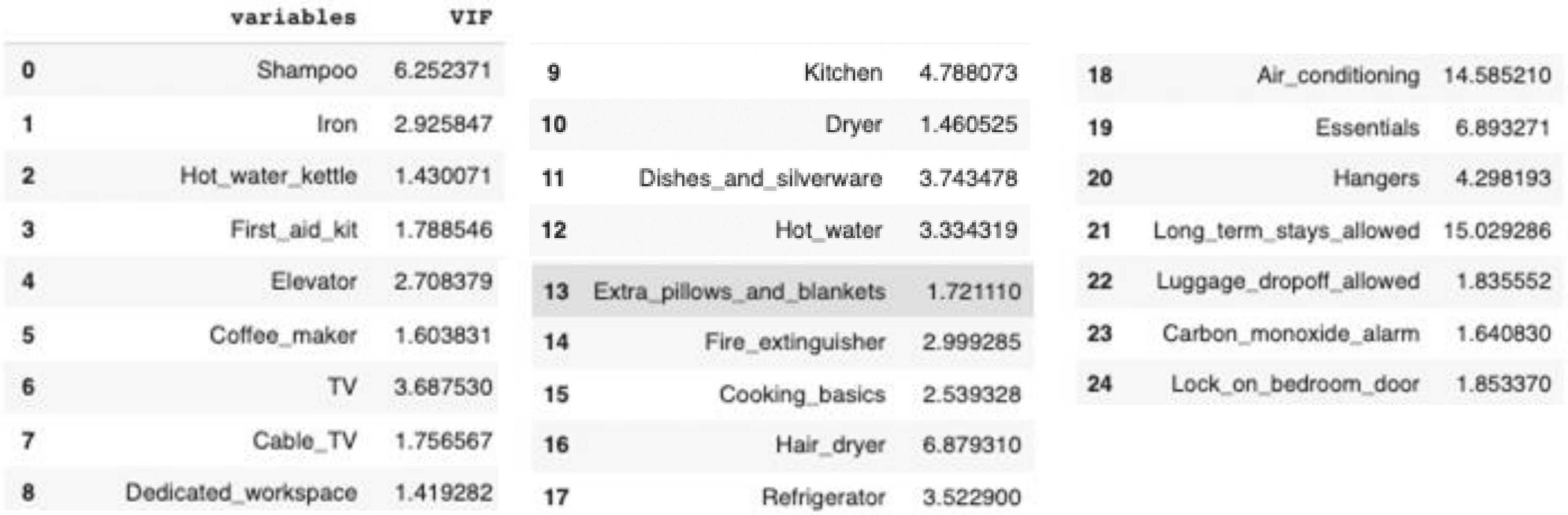
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:      host_is_superhost   No. Observations:                 4050
Model:                            GLM   Df Residuals:                     4039
Model Family:                Binomial   Df Model:                           10
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -932.70
Date:                Sun, 11 Dec 2022   Deviance:                       1865.4
Time:                        18:30:38   Pearson chi2:                 2.81e+03
No. Iterations:                     7
Covariance Type:            nonrobust
===============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                    -2.7300      0.216    -12.648      0.000      -3.153      -2.307
amenities_numbers             0.0698      0.006     10.891      0.000       0.057       0.082
host_acceptance_rate          2.7683      0.242     11.427      0.000       2.294       3.243
host_v_email                 -1.9317      0.196     -9.856      0.000      -2.316      -1.548
neighborhood_overview_exist   1.1165      0.144      7.736      0.000       0.834       1.399
host_identity_verified       -0.8830      0.142     -6.217      0.000      -1.161      -0.605
response_time_a_day          -1.5392      0.300     -5.124      0.000      -2.128      -0.950
response_time_a_few_days     -1.9418      0.597     -3.251      0.001      -3.112      -0.771
beds                         -0.1859      0.050     -3.706      0.000      -0.284      -0.088
instant_bookable             -0.5843      0.147     -3.976      0.000      -0.872      -0.296
host_about_exist             -0.2795      0.136     -2.050      0.040      -0.547      -0.012
===============================================================================================
```

2.2 Models and Results

Model 2: Logistic Regression on Amenity Data

- Taking one step further in looking at the ‘Amenities’.
- Filtered out most generally available in listings in Hong Kong. Picked 35 variables, which accounts for over 80% of the amenities listed in HK.
- Conducted another logistic regression to look at the ones that are most relevant to ‘Superhost’ status?

- Used AIC forward Selection Model and VIF to test for Multicollinearity issues.



	variables	VIF
0	Shampoo	6.252371
1	Iron	2.925847
2	Hot_water_kettle	1.430071
3	First_aid_kit	1.788546
4	Elevator	2.708379
5	Coffee_maker	1.603831
6	TV	3.687530
7	Cable_TV	1.756567
8	Dedicated_workspace	1.419282
9	Kitchen	4.788073
10	Dryer	1.460525
11	Dishes_and_silverware	3.743478
12	Hot_water	3.334319
13	Extra_pillows_and_blankets	1.721110
14	Fire_extinguisher	2.999285
15	Cooking_basics	2.539328
16	Hair_dryer	6.879310
17	Refrigerator	3.522900
18	Air_conditioning	14.585210
19	Essentials	6.893271
20	Hangers	4.298193
21	Long_term_stays_allowed	15.029286
22	Luggage_dropoff_allowed	1.835552
23	Carbon_monoxide_alarm	1.640830
24	Lock_on_bedroom_door	1.853370

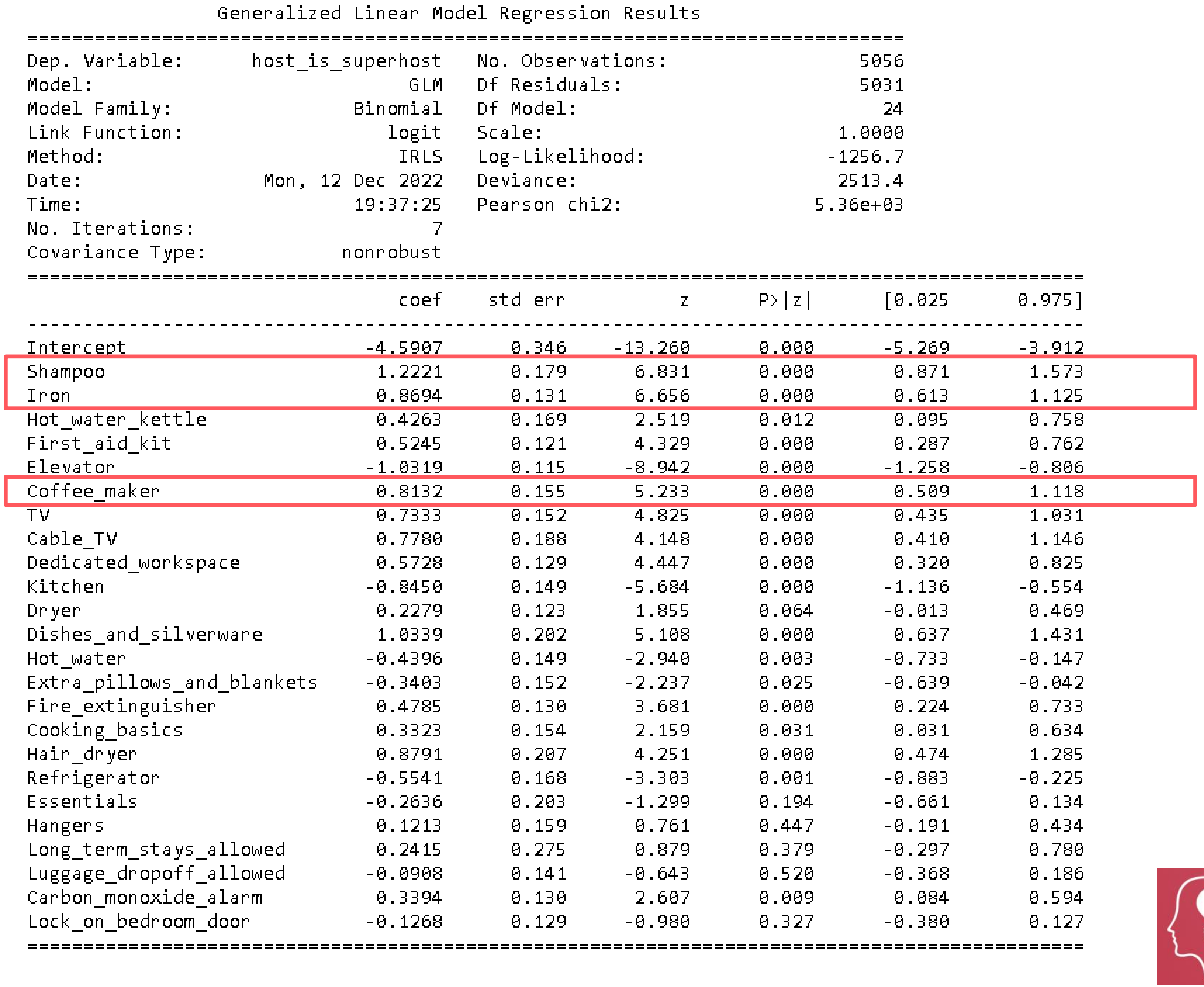
G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

<u>Top 3:</u>

Shampoo

Iron

Coffee maker

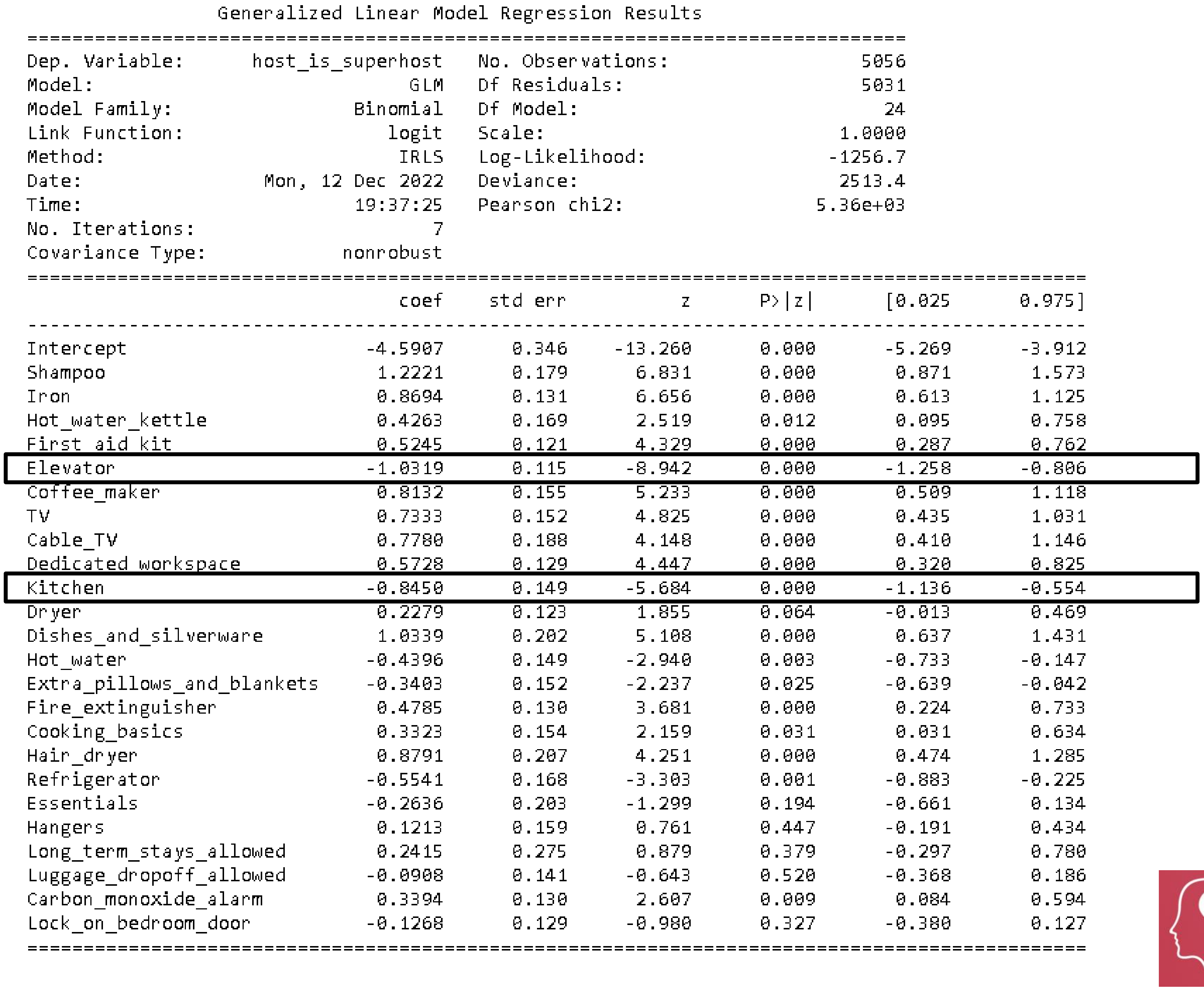


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:      host_is_superhost   No. Observations:                 5056
Model:                            GLM   Df Residuals:                     5031
Model Family:                Binomial   Df Model:                           24
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1256.7
Date:                Mon, 12 Dec 2022   Deviance:                       2513.4
Time:                        19:37:25   Pearson chi2:                 5.36e+03
No. Iterations:                     7
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                     -4.5907      0.346    -13.260      0.000      -5.269      -3.912
Shampoo                        1.2221      0.179      6.831      0.000       0.871       1.573
Iron                           0.8694      0.131      6.656      0.000       0.613       1.125
Hot_water_kettle               0.4263      0.169      2.519      0.012       0.095       0.758
First_aid_kit                  0.5245      0.121      4.329      0.000       0.287       0.762
Elevator                      -1.0319      0.115     -8.942      0.000      -1.258      -0.806
Coffee_maker                   0.8132      0.155      5.233      0.000       0.509       1.118
TV                             0.7333      0.152      4.825      0.000       0.435       1.031
Cable_TV                       0.7780      0.188      4.148      0.000       0.410       1.146
Dedicated_workspace            0.5728      0.129      4.447      0.000       0.320       0.825
Kitchen                       -0.8450      0.149     -5.684      0.000      -1.136      -0.554
Dryer                          0.2279      0.123      1.855      0.064      -0.013       0.469
Dishes_and_silverware          1.0339      0.202      5.108      0.000       0.637       1.431
Hot_water                     -0.4396      0.149     -2.940      0.003      -0.733      -0.147
Extra_pillows_and_blankets    -0.3403      0.152     -2.237      0.025      -0.639      -0.042
Fire_extinguisher              0.4785      0.130      3.681      0.000       0.224       0.733
Cooking_basics                 0.3323      0.154      2.159      0.031       0.031       0.634
Hair_dryer                     0.8791      0.207      4.251      0.000       0.474       1.285
Refrigerator                  -0.5541      0.168     -3.303      0.001      -0.883      -0.225
Essentials                    -0.2636      0.203     -1.299      0.194      -0.661       0.134
Hangers                        0.1213      0.159      0.761      0.447      -0.191       0.434
Long_term_stays_allowed        0.2415      0.275      0.879      0.379      -0.297       0.780
Luggage_dropoff_allowed       -0.0908      0.141     -0.643      0.520      -0.368       0.186
Carbon_monoxide_alarm          0.3394      0.130      2.607      0.009       0.084       0.594
Lock_on_bedroom_door          -0.1268      0.129     -0.980      0.327      -0.380       0.127
================================================================================================
```

Negative Impact : Why?

Kitchen

Elevator



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:      host_is_superhost   No. Observations:                 5056
Model:                            GLM   Df Residuals:                     5031
Model Family:                Binomial   Df Model:                           24
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1256.7
Date:                Mon, 12 Dec 2022   Deviance:                       2513.4
Time:                        19:37:25   Pearson chi2:                 5.36e+03
No. Iterations:                     7
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                      -4.5907      0.346    -13.260      0.000      -5.269      -3.912
Shampoo                         1.2221      0.179      6.831      0.000       0.871       1.573
Iron                            0.8694      0.131      6.656      0.000       0.613       1.125
Hot_water_kettle                0.4263      0.169      2.519      0.012       0.095       0.758
First_aid_kit                   0.5245      0.121      4.329      0.000       0.287       0.762
Elevator                       -1.0319      0.115     -8.942      0.000      -1.258      -0.806
Coffee_maker                    0.8132      0.155      5.233      0.000       0.509       1.118
TV                              0.7333      0.152      4.825      0.000       0.435       1.031
Cable_TV                        0.7780      0.188      4.148      0.000       0.410       1.146
Dedicated_workspace             0.5728      0.129      4.447      0.000       0.320       0.825
Kitchen                        -0.8450      0.149     -5.684      0.000      -1.136      -0.554
Dryer                           0.2279      0.123      1.855      0.064      -0.013       0.469
Dishes_and_silverware           1.0339      0.202      5.108      0.000       0.637       1.431
Hot_water                      -0.4396      0.149     -2.940      0.003      -0.733      -0.147
Extra_pillows_and_blankets     -0.3403      0.152     -2.237      0.025      -0.639      -0.042
Fire_extinguisher               0.4785      0.130      3.681      0.000       0.224       0.733
Cooking_basics                  0.3323      0.154      2.159      0.031       0.031       0.634
Hair_dryer                      0.8791      0.207      4.251      0.000       0.474       1.285
Refrigerator                   -0.5541      0.168     -3.303      0.001      -0.883      -0.225
Essentials                     -0.2636      0.203     -1.299      0.194      -0.661       0.134
Hangers                         0.1213      0.159      0.761      0.447      -0.191       0.434
Long_term_stays_allowed         0.2415      0.275      0.879      0.379      -0.297       0.780
Luggage_dropoff_allowed        -0.0908      0.141     -0.643      0.520      -0.368       0.186
Carbon_monoxide_alarm           0.3394      0.130      2.607      0.009       0.084       0.594
Lock_on_bedroom_door           -0.1268      0.129     -0.980      0.327      -0.380       0.127
===============================================================================================
```

G18 CONSULTING

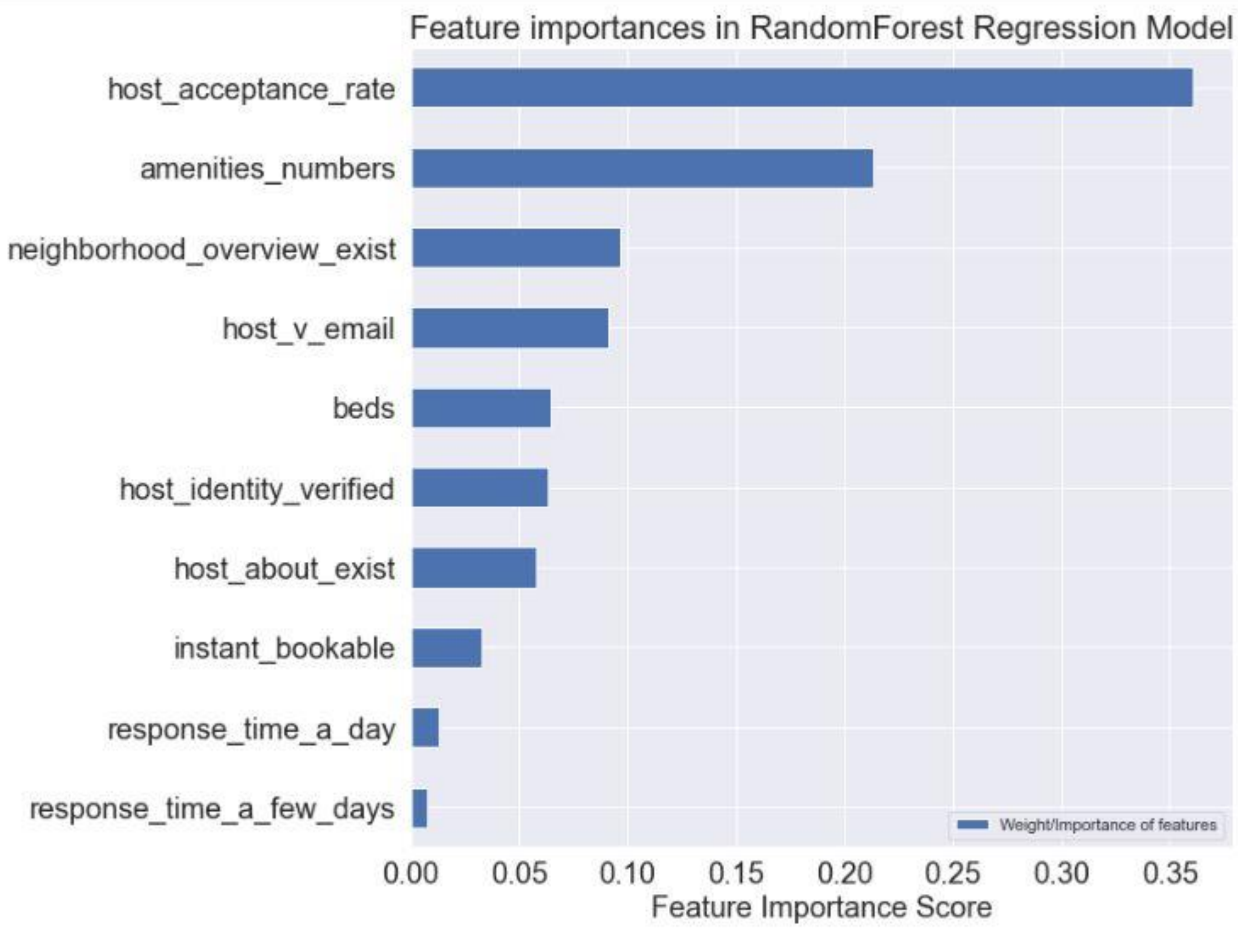
WE EAT YOUR PROBLEM FOR BREAKFAST

2.2 Models and Results

Model 3 : Random Forest Classification on Feature Data

- Random Forest Regression on the variables identified for the final Model 1 above to verify the importance of each variable from another perspective.

- Both models produce similar results in that the number of available amenities and the hosts' acceptance rate as well as the existence of neighbourhood overview have the highest impact on becoming "Superhosts".



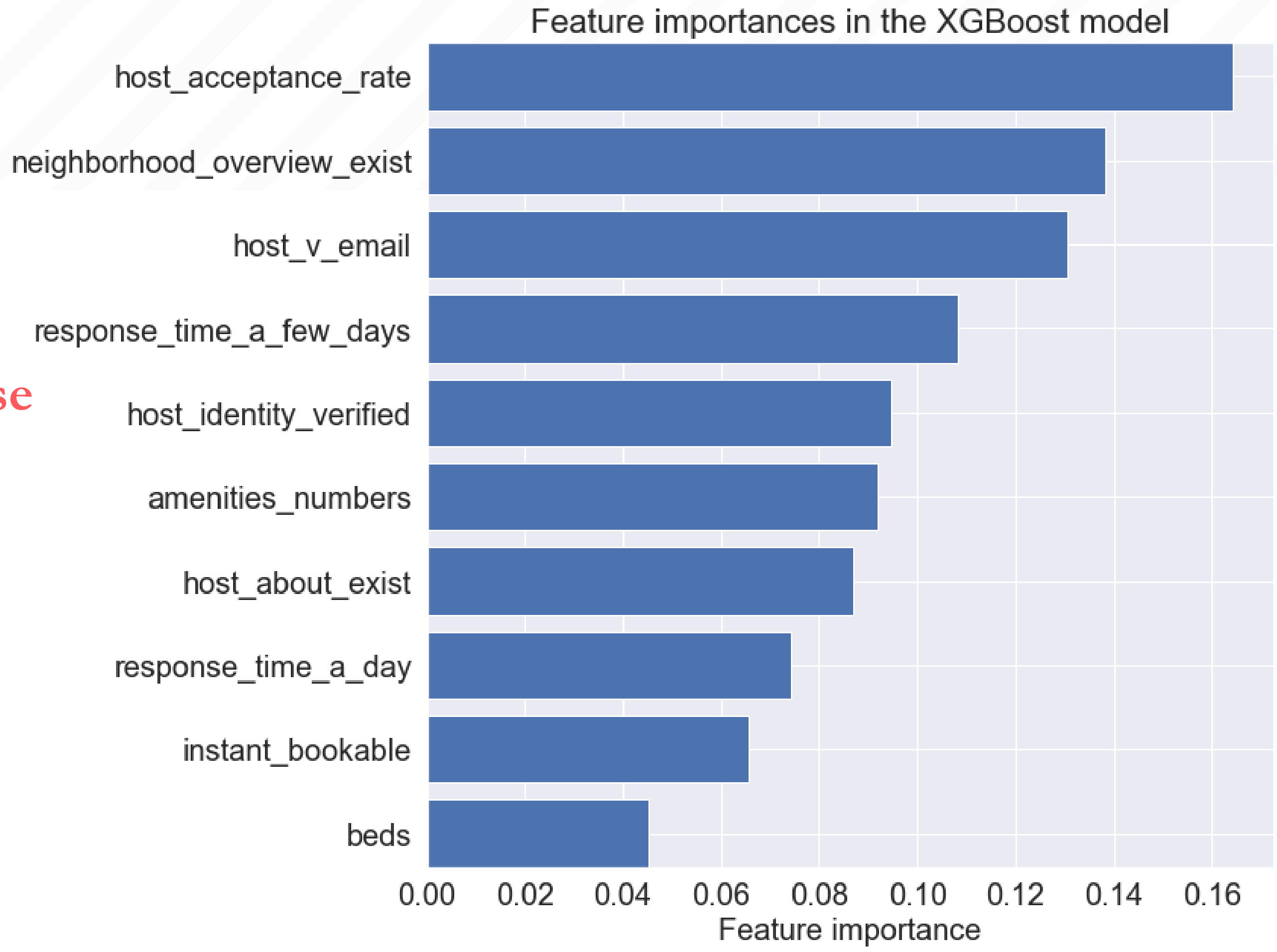
2.2 Models and Results

Model 4: XGBoost

- XGBoost model to verify the importance of each variable identified in the final Model 1

- Hosts having **verification emails** and **slow response time** have the highest importance score.

- we must be cautious that this model does not tell whether a feature positively or negatively impact the prospective of becoming "Superhosts".



G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

3. Recommendations

- Recommendations for hosts

3. Recommendations

Airbnb should encourage Hong Kong hosts to …

Respond to clients within a day

Accept bookings as much as possible

Remove listing for future dates in advance if they cannot accept booking

Include overview of the neighborhood in the listing

Provide more amenities. Not just the staples but extra touches for your guests

Top amenities:
Iron, TV, equipped kitchen, equipped bathroom, workspace, safety equipment

Thanks!

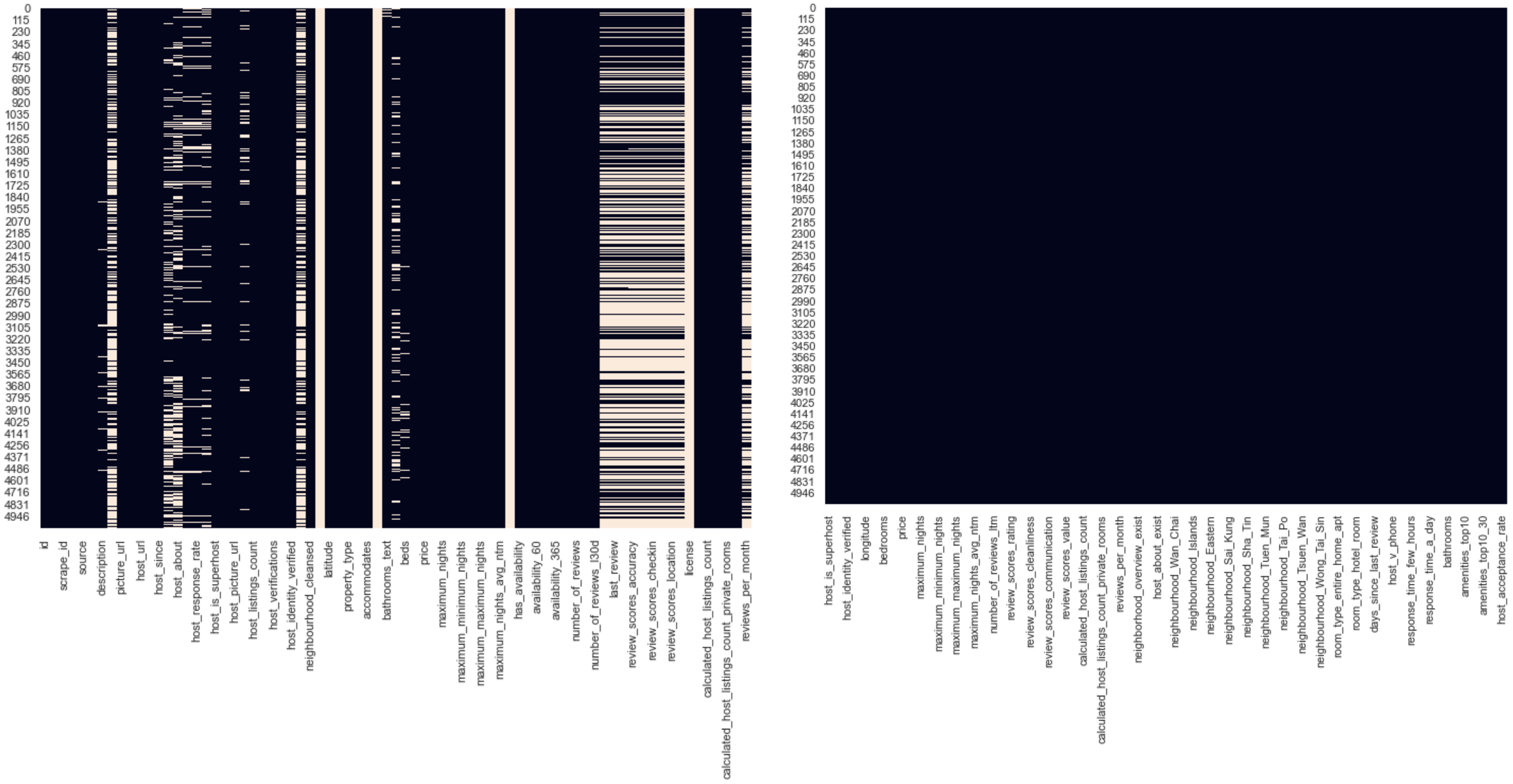
Any questions?

You can find us at

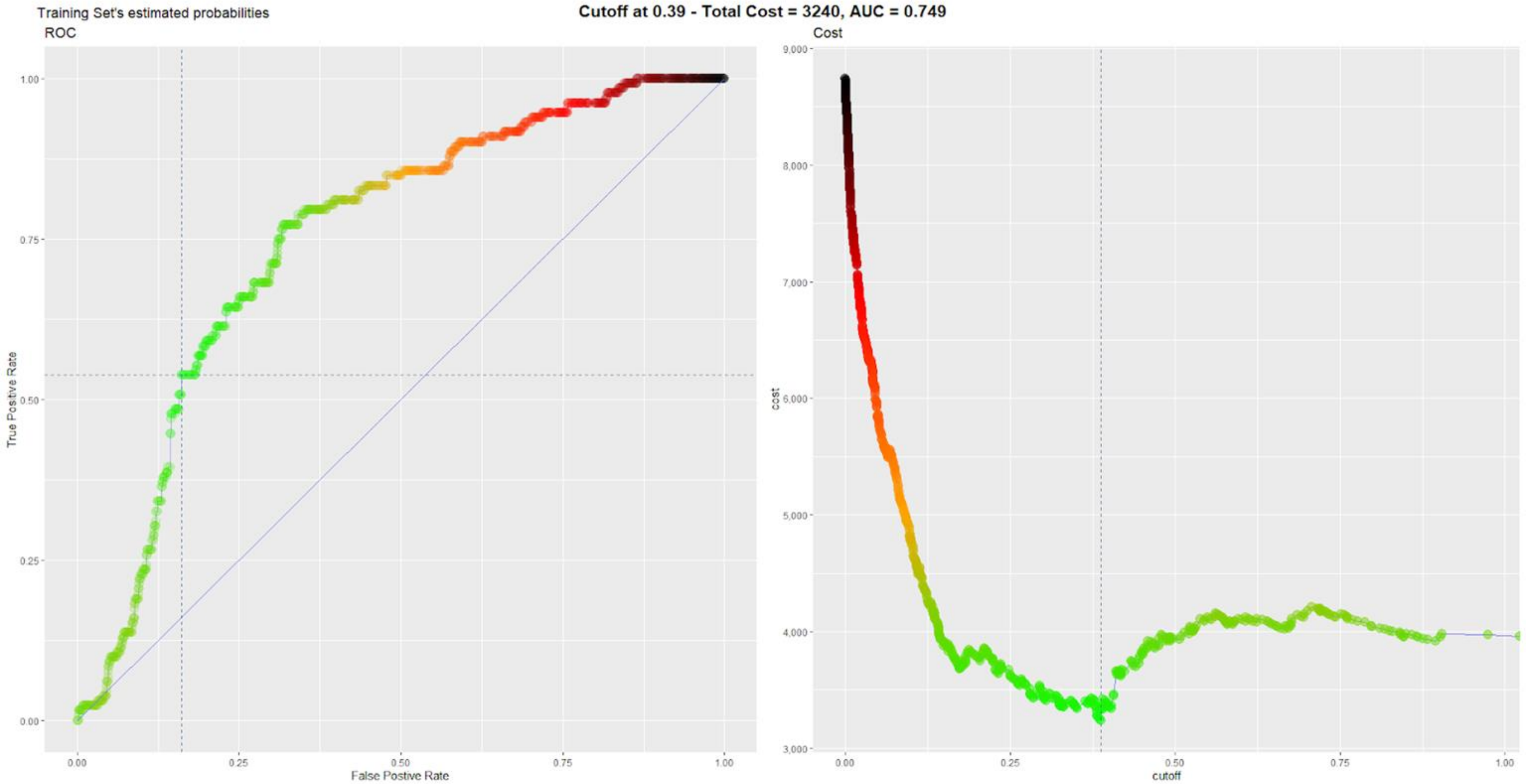
✉ ProjectMarvel@G18Consulting.co

4.
Appendix

- Data Cleaning Process
- Classification in respect of Logistic Regression
- Model Evaluation Metrics



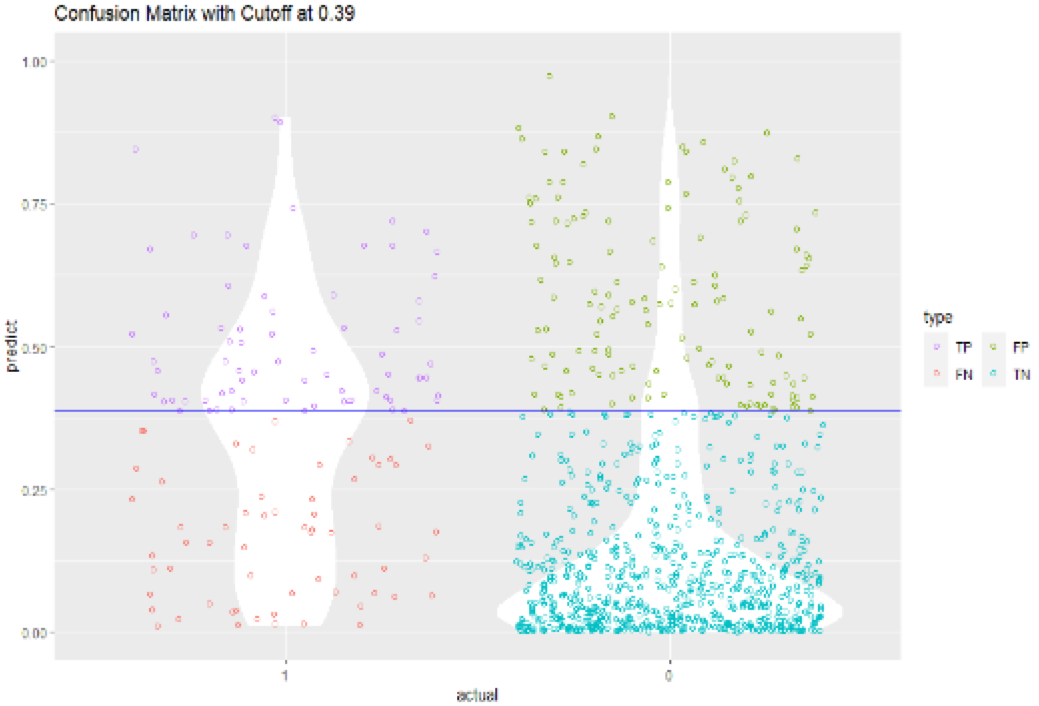
Shows the plot of missing data before and after clean up.



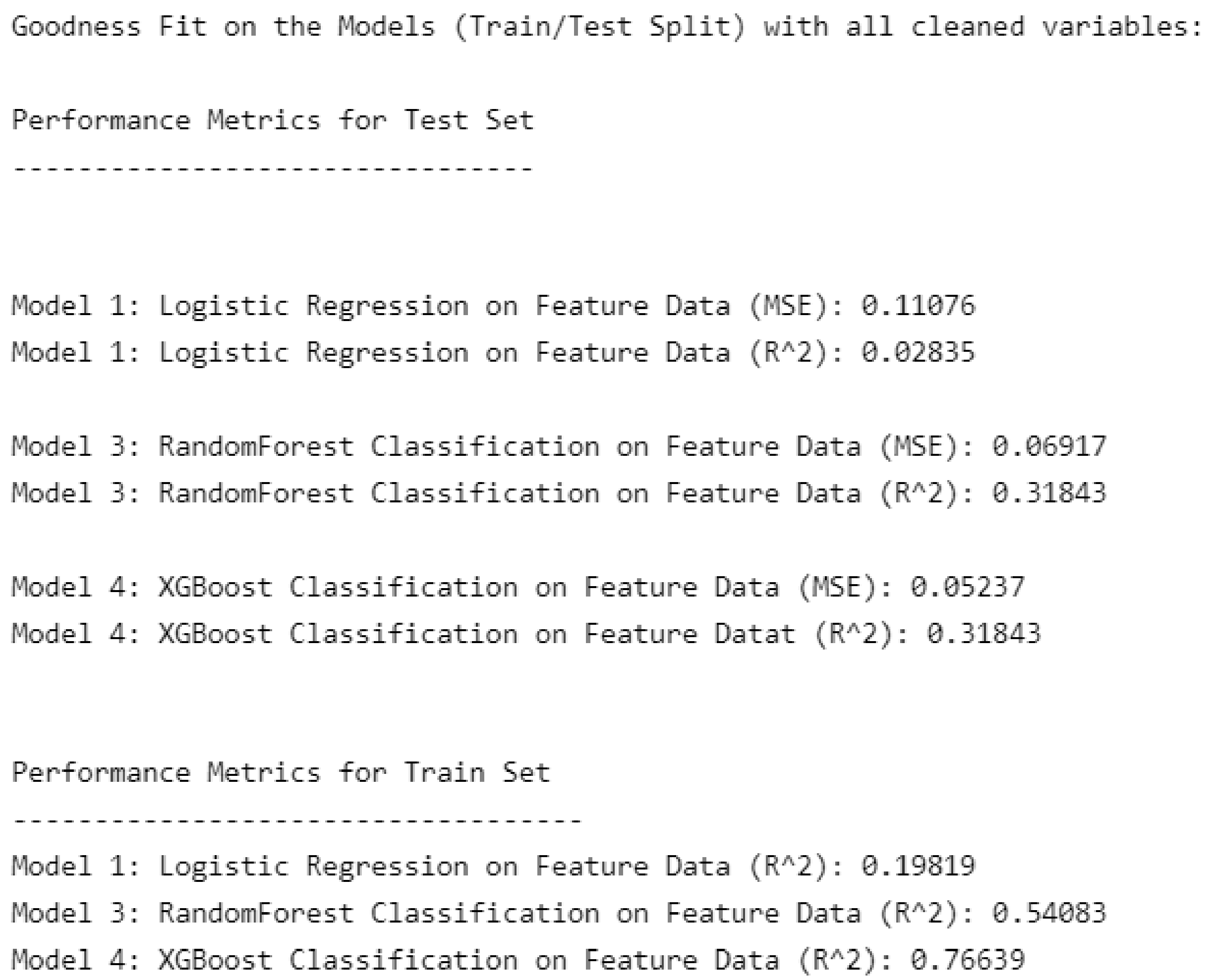
Based on the 5-folded cross-validation approach together with an assumption that the cost of wrongly predicting the hosts to become a Superhost (false negative) is triple the cost of not predicting some Superhost in advance (false positive), we determine the optimal cut-off point to be 0.39, which could minimise the total cost. That will lead to an AUC of 0.749.

	FALSE	TRUE
0	733	141
1	65	67

[1] The sensitivity is 0.5379
[2] The specificity is 0.8387
[3] The Misclassification Rate is 0.2008



It could be concluded that the model gives reasonable prediction accuracy with the TP rate (Sensitivity) of 53.79%, whereas the overall Misclassification Rate is 20.08%.



```
Goodness Fit on the Models (Train/Test Split) with all cleaned variables:

Performance Metrics for Test Set
--------------------------------



Model 1: Logistic Regression on Feature Data (MSE): 0.11076
Model 1: Logistic Regression on Feature Data (R^2): 0.02835

Model 3: RandomForest Classification on Feature Data (MSE): 0.06917
Model 3: RandomForest Classification on Feature Data (R^2): 0.31843

Model 4: XGBoost Classification on Feature Data (MSE): 0.05237
Model 4: XGBoost Classification on Feature Datat (R^2): 0.31843


Performance Metrics for Train Set
-----------------------------------
Model 1: Logistic Regression on Feature Data (R^2): 0.19819
Model 3: RandomForest Classification on Feature Data (R^2): 0.54083
Model 4: XGBoost Classification on Feature Data (R^2): 0.76639
```

Using MSE and R^2, we have evaluated each models we used in this report.
XGBoost Classification shows the highest R^2, and the lowest MSE.