Exploratory Data Analysis on Raw Data

Reviewing the dataset

1. 1284 rows or observations of Airbnb listings and 19 variables.
2. No missing values.
3. No duplicate values found, checked with remove duplicate function in excel.

EDA

- These are a few insights we got from the data visuals on raw data. The average price is higher when it's an entire home or apartment. (Figure 1)
- The average price is higher for when a host is rated as a super host. (Figure 2)
- There is no apparent trend in average price against guest satisfaction rating. (Figure 3)
- The average price follows an increasing trend when it comes to cleanliness rating. (Figure 4)
- The scatterplots between price and city and metro distance show some extreme values when the distance is closer. (Figure 5 and 6)
- The average price increases till the number of rooms is 4 then it starts decreasing. (Figure 7)
- There are some extreme values and outliers in the realSum observations detected by box plot, scatter plot and histogram and the descriptive analysis shows that the mean is around 240 with standard deviation around 230. (Figure 8, 9, 10, 11)
- We've also come up with a correlation matrix before data cleaning (CM1)

Data Cleaning (Round 1)

Changing Data Values

Top/Bottom 2.5% of the data i.e., 71 observations of real_Sum replaced with median to remove extreme values as shown in Histogram. (Figure 12) Mean and Standard Deviation after replacing with median is around 221 and 104 respectively. There is a significant change in standard deviation. (Figure 13)

Data Deletion

Outliers from Categorical Variables data removed, and 1250 observations left. Extreme values – Top/Bottom 2.5% data of numeric variables i.e., dist, metro-dist, att_index_normalized, rest_index_normalized. 1098 observations left after deletion.

Data Imputation

There were no NA values in any variable of Berlin's data.

EDA (Round 2)

Univariate Analysis/Examining the Target Variable

From the summary statistics (SM1) (mean > median) and Cullen and Frey graph (CF1), it seems the prices have a right skew. The observations seem to be close to the gamma line hence indicating it can best be approximated with a Gamma distribution.

Fitting To the Different Distributions:

1. **Gaussian Distribution:** as DP(1) shows, normal distribution isn't a good approximation for Airbnb prices.
2. **Exponential Distribution**: As DP(2) shows, an even poor fit.

3. **Gamma Distribution:** As DP (3) shows, price seems to best be modeled by a gamma distribution.
4. **Weibull Distribution:** as DP (4) shows, this is the second best fit.

Here are the AICs of all 4 fits, respectively: 13247.53, 14035.48, 12873.47, 13087.23.

Data Cleaning (Round 2)

- After making Histograms for all variables on R (HT1, HT2, HT3, HT4, HT5, HT6, HT7, HT8, HT9, HT10, HT11, HT12, and HT13) some extreme values and outliers were removed from the following variables.
- 2 or higher metro distance has been removed.
- Distance less than 1 and higher than 13 has been removed. 13 observations deleted.
- Rest index norm higher than 70 removed. 19 observations removed.
- Attr_index 40 or higher removed. 11 observations removed. Less than 5 too. 11 observations removed. 22 observations removed in total.
- The Real Sum's top 2% was deleted via conditional formatting (CND) and all values below $100.
- Same has been done for all other variables. Cleanliness rating of 4, 6, and 7. 37 observations removed.
- Guest sat below 80 also removed. 8 observations eliminated.
- Bedrooms $\geq$3 have been removed; 22 observations deleted.
- Room_shared column deleted because all of them were not shared. False for all.
- Remove non normal columns of attr-index and rest_index.
- Moreover, all outliers have been removed: the bottom 2.5% and top 2.5% of data of each variable has been removed.

Exploratory Data Analysis Post Data Cleaning

Reviewing the Data Set

951 rows or Airbnb listings with 16 variables.

Univariate Analysis

Examining the distribution of the target variable using descriptive statistics (SM2) and visualizing the target variable's distribution using histogram (YHT) and density plot plus Cullen and Frey Graph for goodness of fit. (CF2)

Fitting to Gaussian Distribution (DP5)

Fitting Exponential Distribution (DP6)

Fitting Gamma Distribution (DP7): this one is again the best fit as the 4 plots show a close match.

Numerical Variables Analysis

(See Appendix (N1, N2)**)**

Demand can be in the negative too. The only negative valued variable in our data. Externally sourced.

Dummy and Categorical Variables Analysis

(See Appendix (DV)**)**

Bivariate Analysis

For accessing the relationship between the target variable and each independent variable. The variables which seem useless from the visual bivariate analysis are: Demand, Attractions Count, Crime Rate, Multi, and Host is Superhost. (BV1, 2, 3,...17)

Correlation Analysis

(See Appendix (CA1, CA2))

P_values

(See Appendix (P1, P2))

Correlogram or Heatmap

From the matrix (CM2), we can see that some variables have almost no impact on realSum or price of an Airbnb. These include Host_Superhost, cleanliness_Rating, guest_satisfaction, and Demand.

Checking Multicollinearity

1. Private room seems to be somewhat multicollinear with person_capacity, negatively though
2. Person_Capacity with bedrooms, but not too significant.
3. Guest_sat and cleanliness_rating seem highly correlated
4. Dist and attr_index_norm and rest_index_norm, negatively.
5. Metro_dist and rest_index_norm
6. Rest_index and attr_index high multicollinearity.
7. Crime_rate with attr_index_norm,
8. Crime_rate and Attractions_count
9. Crime_rate and Demand (CM2)

Data Transformations

From the scatterplots (BV1…BV17), it doesn't seem necessary to do any data transformations right now, but we can't really give a final word until we have made our model and tested the linear regression assumptions.

Feature Selection and Engineering

Using the Lat and Ion variables, I tried coming up with a new variable called Haversine for each Airbnb listing. Here's the formula for spherical cosines:

cellx=SIN((lat2 – lat1)/2)^2+COS(lat1)*COS(lat2)*SIN((lon2-lon1)/2)^2

celly=2*Atan2(sqrt(1-cellx),sqrt(cellx)) // big trick here is that ATAN2

Model Selection

Since our Y is continuous, we have used a linear regression. Here are the regression results for the entire dataset before doing any train/test split. We have done this in order to test if our data meets all the assumptions of linear regression. Here's a summary of the original linear model (MS 1). And here's a summary of the stepwise regression model.

Linearity

There exists a linear relationship between the explanatory variables and the target variable, Airbnb listings. As we can see from the residuals vs Fitted plot (MS 2), there is a slight pattern in the residuals plot. So, linearity might not be met here.

Autocorrelation

H0 (null hypothesis): There is no correlation among the residuals.

Ha (alternative hypothesis): The residuals are autocorrelated.

Since the p-value is 0.77 (MS 3), which is much higher than 0.05 so we can easily reject the alternative hypothesis and say there's no autocorrelation between the residuals in our model.

Homoscedasticity

We can check this assumption using the Scale-Location plot. In this plot we can see the fitted values vs the square root of the standardized residuals. Ideally, we would want to see the residual points equally spread around the red line, which would indicate constant variance. In the (MS 4) plot, we can see that the residual points are not all equally spread out. Thus, this assumption is not met. We can also use the Non-Constant Error Variance (NVC) Test using R's built in function called nvcTest to check this assumption. This will output a p-value which will help you determine whether your model follows the assumption or not. The null hypothesis states that there is constant variance. Thus, if you get a p-value> 0.05, you will fail to reject the null. But our p value is well below 0.05 (MS 5). So, it means we can reject H0 or null and conclude there is variance in the error term. One common solution to this problem is to calculate the log or square root transformation of the outcome variable.

Normality of Residuals

The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. But as we can see from our QQ plot (MS 6), the residuals don't follow a normal distribution since the plot doesn't have a straight line.

Zero Conditional Mean

This is the assumption that the mean values of our error term are around 0. We will again use the Residuals vs Fitted plot and would ideally want to see the red line flat on 0, which would indicate that the residual errors have a mean value of zero. But as our plot shows (MS 7), the mean values for higher fitted values or higher realSum values of Airbnb listings isn't 0. Far from it.

Little Outliers and High Leverage Points

From our plot (MS 8), we can see that there are 3 extreme points: 170, 617, and 630. But it is just 3 out of 951 residuals. That checks out. However, there are 4 high leverage points whose leverage exceeds 0.05 on the x-axis. But again, it's just 4 points out of 951, so that checks out too!

Data Transformation and Testing Assumptions Again

Since most of the linear regression assumptions have failed, we need to transform our data to make it linear.

Logarithmize the dependent variable

This transformation alone has made huge improvements in residuals vs fitted i.e. linearity, and in QQ plot, normality of residuals. (DT 1 and 2) R^2 has already increased by almost 5% (DT 3).

Logarithmize Guest_satisfaction

R^2 has now fallen to 85% and there's no noticeable improvements in meeting the linear regression assumptions as shown here (DT 4). So, we shall not use this transformation.

Logarithmize attr_index and rest_index

R^2 is still not the most optimal (DT 5), which was 0.88 or 88% in stage 1. And no improvements in meeting LR assumptions either (DT 6), so we won't apply this transformation either. Final model has a log(realSum) and everything else is just the same.

Model Training and Visualization

The dataset has been split into a train and test set, with the former having 600 observations and the latter having 351 observations. The model has an R squared value of 89.6% (MV 1). It is incredibly good at explaining the variation in the SST of the price of an Airbnb listing. The residual sum of errors (RSE) is just 0.118. Since the p value of the F-statistic is less than 0.05, we can conclude the explanatory variables are jointly significant so need not be changed. And this (MV 2) is summary of the stepwise regression, which has also seen improvement in its accuracy after the data transformation. And the R square figure for the stepwise model is slightly higher and the RSE is a bit lower than the previous OLS model. We will use this for our interpretation since it contains only the significant variables.

Model Interpretation

After the stepwise regression, only the most significant variables remain while others are eliminated. Notice, however, R squared value has not changed going opposite to our expectations (MV 2). Analyzing the coefficients, the variable that impacts the prices of Airbnb the most is cleanliness rating. The second highest impact is of person capacity.

Model Deployment, Reporting, and Comparison

The snapshot of the OLS model (OLS 1) shows how this model performs on each KPI. We already know its R^2 is quite high, 89.6%. But we have to evaluate how well it generalizes. We need to test it on the test set. The first section shows the model's performance on the overall data. Its Mape, mdape, and MSE are all fairly low, hence the model gives us very little errors. It gives about 1.6% error only, averaging all 3 KPIs. But on the test set it does perform comparatively poorly. It is still giving around only 2% of error, but that's a bit higher than 1.6% on the overall dataset. The model might be overfitting a bit. And the snapshot (ST 1) shows just the same analysis on the overall and test set but for stepwise regression. Although stepwise's regression has a slightly higher R^2, it performs much worse on our KPIs than the OLS model does. It gives a 9.82% error if we look at MAPE, and on the test set it gets even worse, an 11.6% error. Its RSE is about the same, but we must consider unbiased KPIs such as MAPE. This model also overfits, more so than the OLS one. Hence the R-squared value is high this means that the variables we have chosen for the model explains variation to a higher degree in the prices.

Feature Importance Analysis

Decision Tree Model

Decision Tree Analysis for Berlin

Model_1, model_2, and model_3 are for Train data and all three models have the same deviance value, indicating similar overall performance in terms of fitting the data. However, we can consider the complexity of the trees to make a decision.

Model_1 has the highest number of terminal nodes (15), indicating a more complex tree structure. Model_2 has fewer terminal nodes (8) with minbucket = 25, while Model_3 has the fewest terminal nodes (5) with minbucket = 50. Generally, a simpler tree is preferred because it tends to be more interpretable and less prone to overfitting. Based on the complexity of the trees, Model_3 appears to be the better choice among the three options.

Like train data, test data too has the same deviance values for model_4, model_5, and model_6, but we will choose model_6 as it has the lowest number of nodes (8) with a minbucket = 20 and a simpler tree structure.

In the Train data model, the Private_Room is significant in determining the outcome, as it is used as the first split in the decision tree. The variable guest_satisfaction_overall also seems to be significant as it is used as a split in both the Private_Room=1 and Private_Room=0 groups. The variable cleanliness_rating is used as a split only within the Private_Room=1 group, indicating it is significant in that particular group.

In the Test data model, Private_Room is significant as it is used as the first split in the decision tree. The variable "guest_satisfaction_overall" is used as a split in both the Private_Room=1 and Private_Room=0 groups, suggesting its significance in predicting the outcome. Within each group, there are further splits based on "guest_satisfaction_overall," indicating that the variable has predictive power within those subgroups. However, the cleanliness_rating is not relatively a significant variable in test data.

Business Insights

There are several potential parties who could benefit from using or purchasing our model:

1. Airbnb Hosts:

   - Airbnb hosts can use the model to set appropriate pricing for listings, maximizing the earnings they can make while remaining competitive.

   - The model can help them optimize their pricing based on factors such as location, property characteristics, amenities, and ratings.

2. Real Estate Investors:

   - Real estate investors interested in purchasing properties for short-run rental purposes can utilize our model to determine the profitability of different properties.

   - The model can help investors estimate the potential rental income they can earn, hence helping them make informed investment decisions.

3. Airbnb Management Companies:

   - We can sell our model to companies or individuals providing property management services for Airbnb listings.

   - The model can assist in optimizing two things— occupancy rate and rental income—for the managed properties, benefiting the management company and the property owners.

4. Pricing Analytics Platforms:

- Pricing analytics platforms specializing in offering tools that optimize price in the short-term rental industry can cash out a lot for our model.

- Our model can boost the accuracy of their pricing algorithms, providing added value to their customers and raising profits.

5. Research Institutions or Academia:

- Researchers focusing on what impacts the prices in Airbnb and rental market can find our regression model useful.

- The model can aid their studies and simulations, helping them come up with insights into market dynamics or consumer behavior.

6. Data-Driven Consultancy Firms:

- We can sell our model to data-driven consultancy firms specializing in providing analytical insights to businesses in the travel and hospitality sector.

- They can integrate our model into their consulting services, ensuring their clients have good pricing guidance to optimize profitability.

Appendix

Explanatory Data Analysis on Raw Data

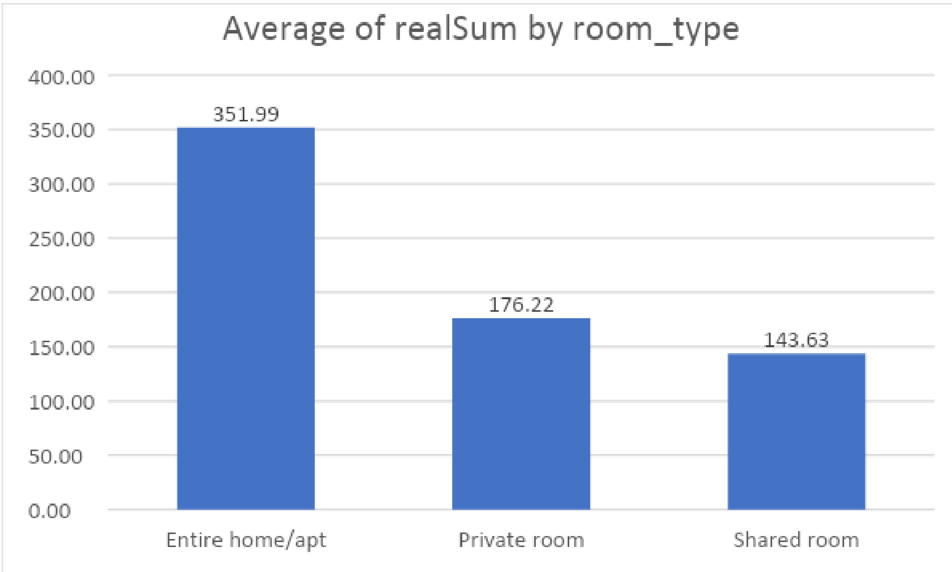


Figure 1

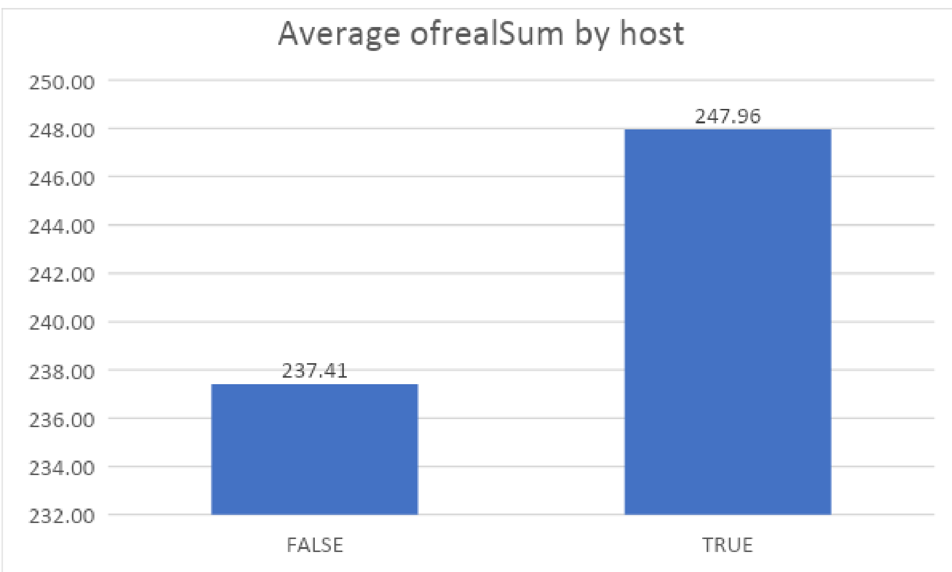


Figure 2

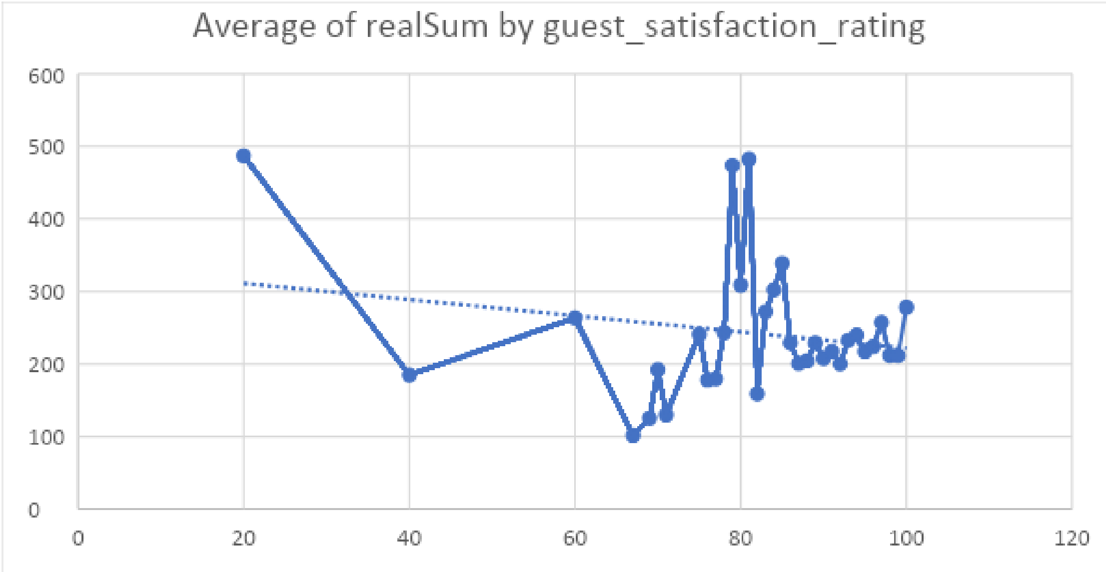


Figure 3

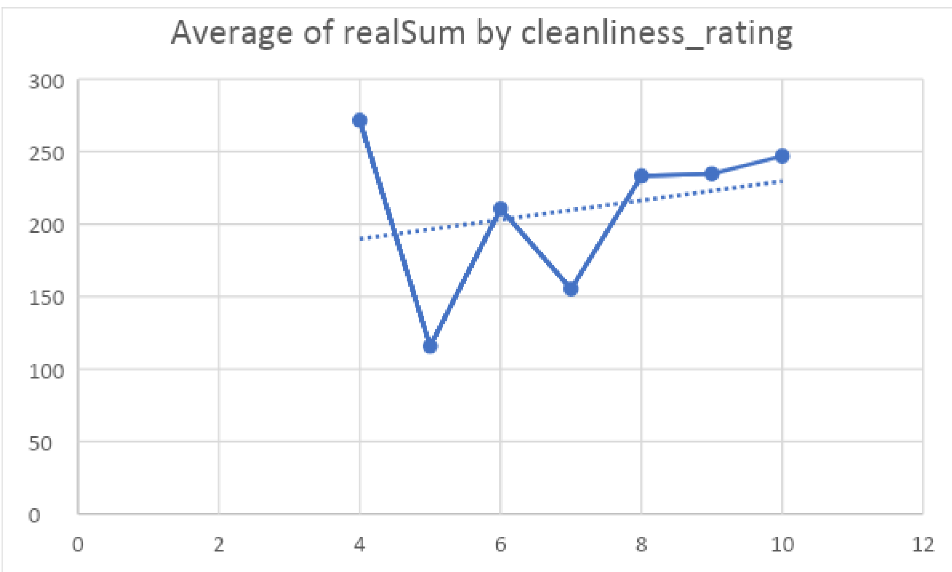


Figure 4

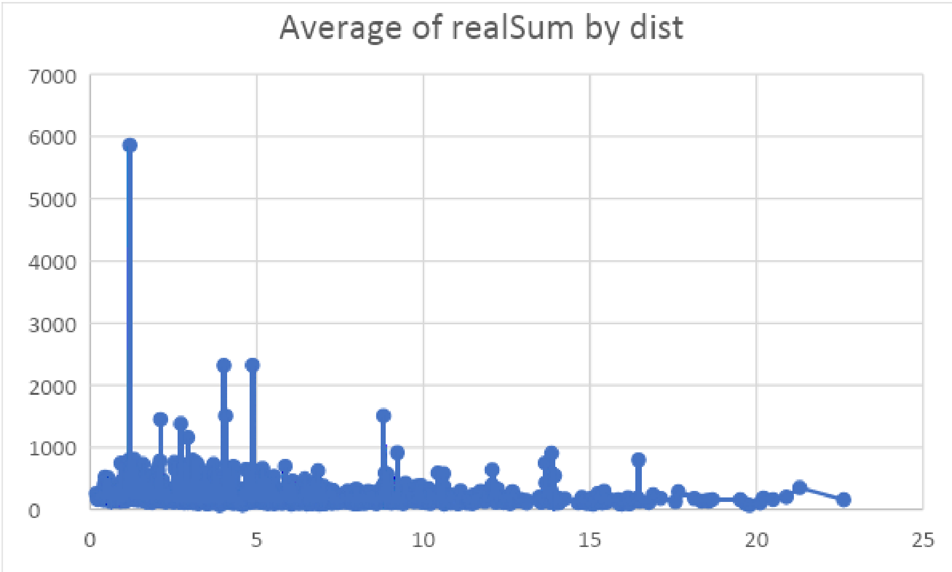


Figure 5

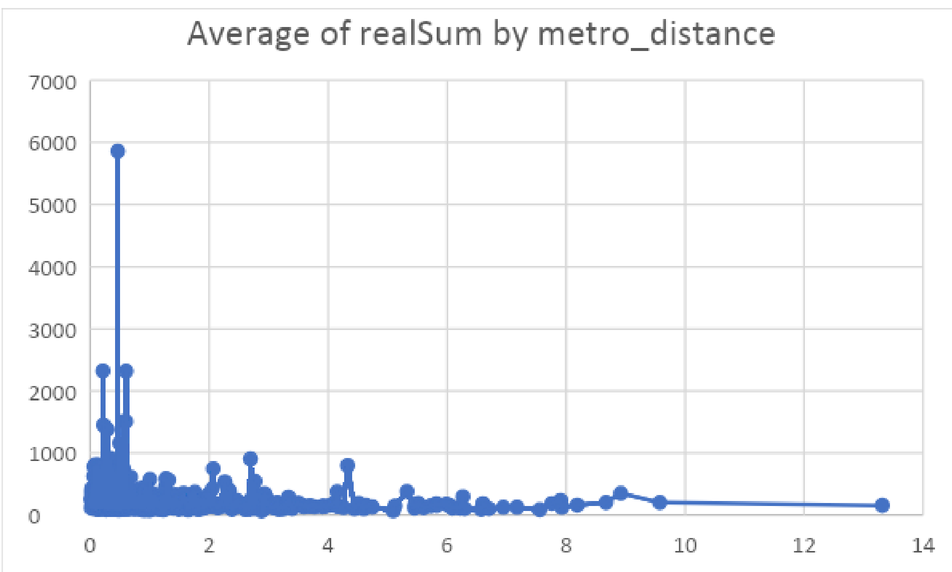


Figure 6

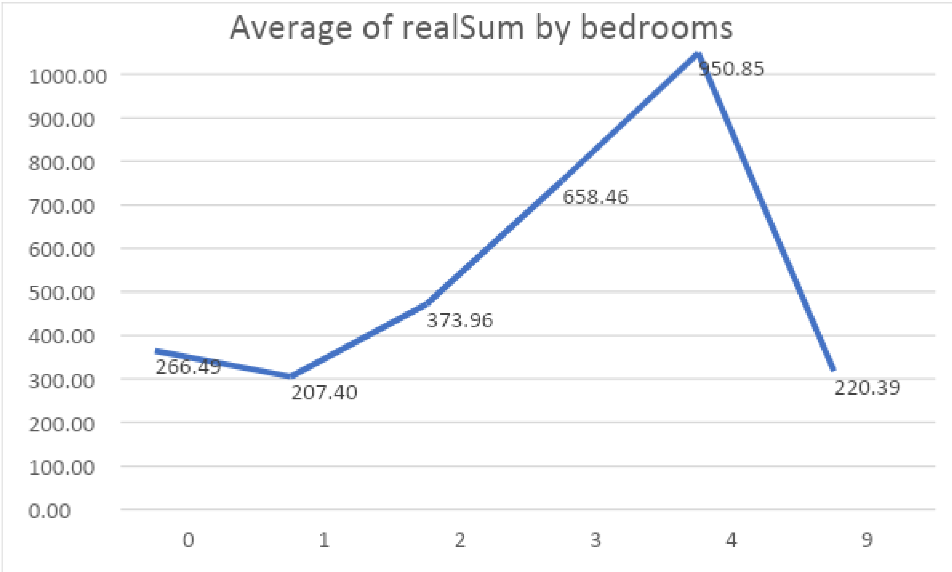


Figure 7

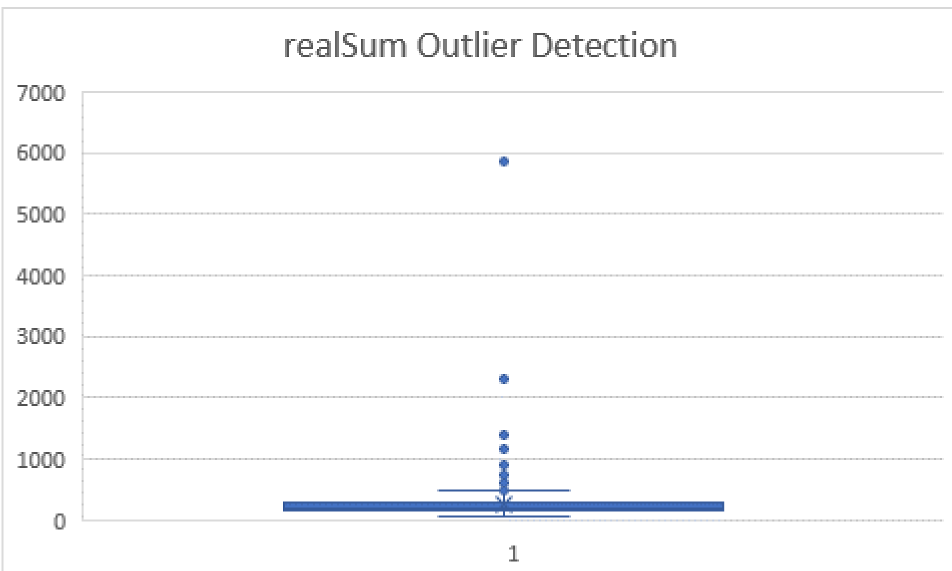


Figure 8

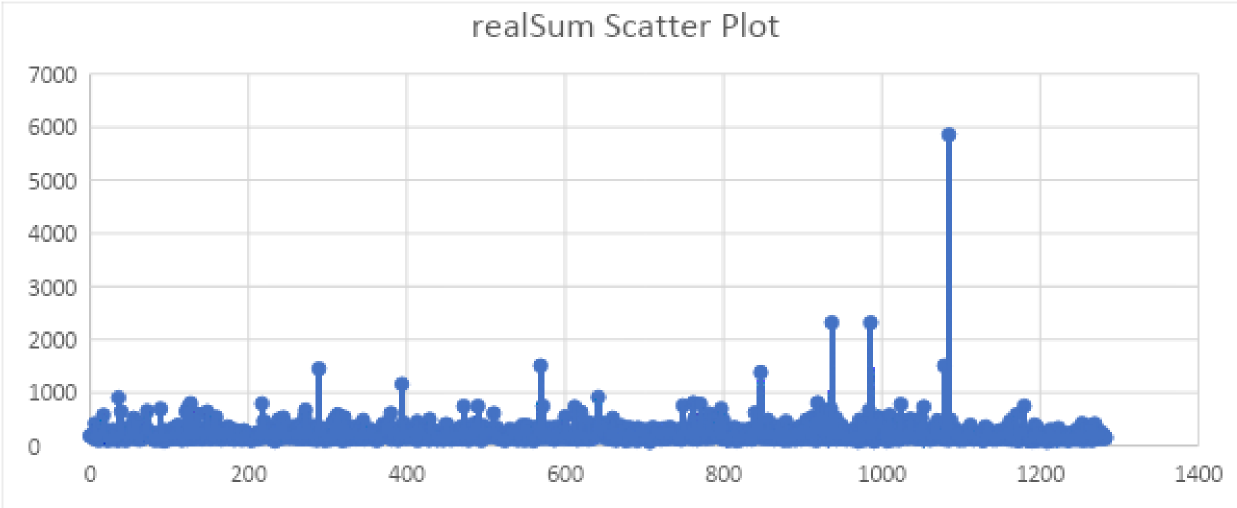


Figure 9

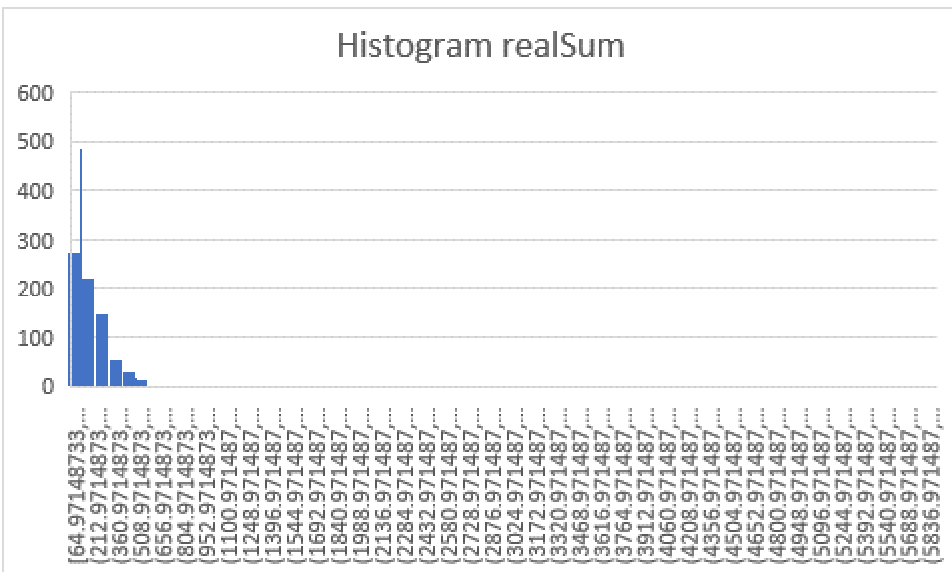


Figure 10

realSum	
Mean	240.2204
Standard Error	6.427553
Median	187.7863
Mode	150.7432
Standard Deviation	230.3182
Sample Variance	53046.46
Kurtosis	287.8197
Skewness	13.27989
Range	5792.512
Minimum	64.97149
Maximum	5857.483
Sum	308443
Count	1284

Figure 11

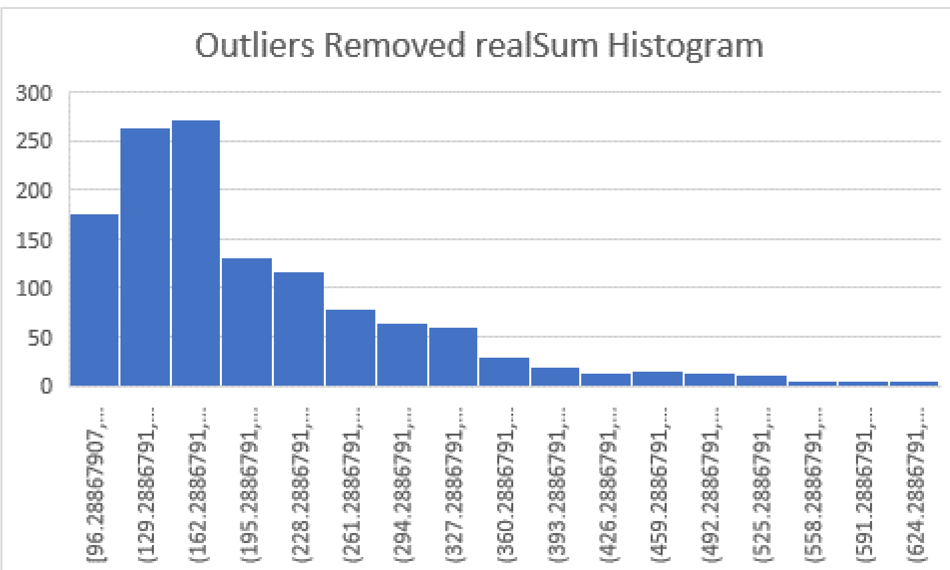


Figure 12

Values	
Average of realSum	221.6577451
StdDev of realSum4	104.8881155
Max of realSum3	644.8069552
Min of realSum2	96.28867907

Figure 13

Summary statistics from descdist(Berlin$realSum, discrete = FALSE)

min: 96.28868   max: 644.807

median: 187.7863

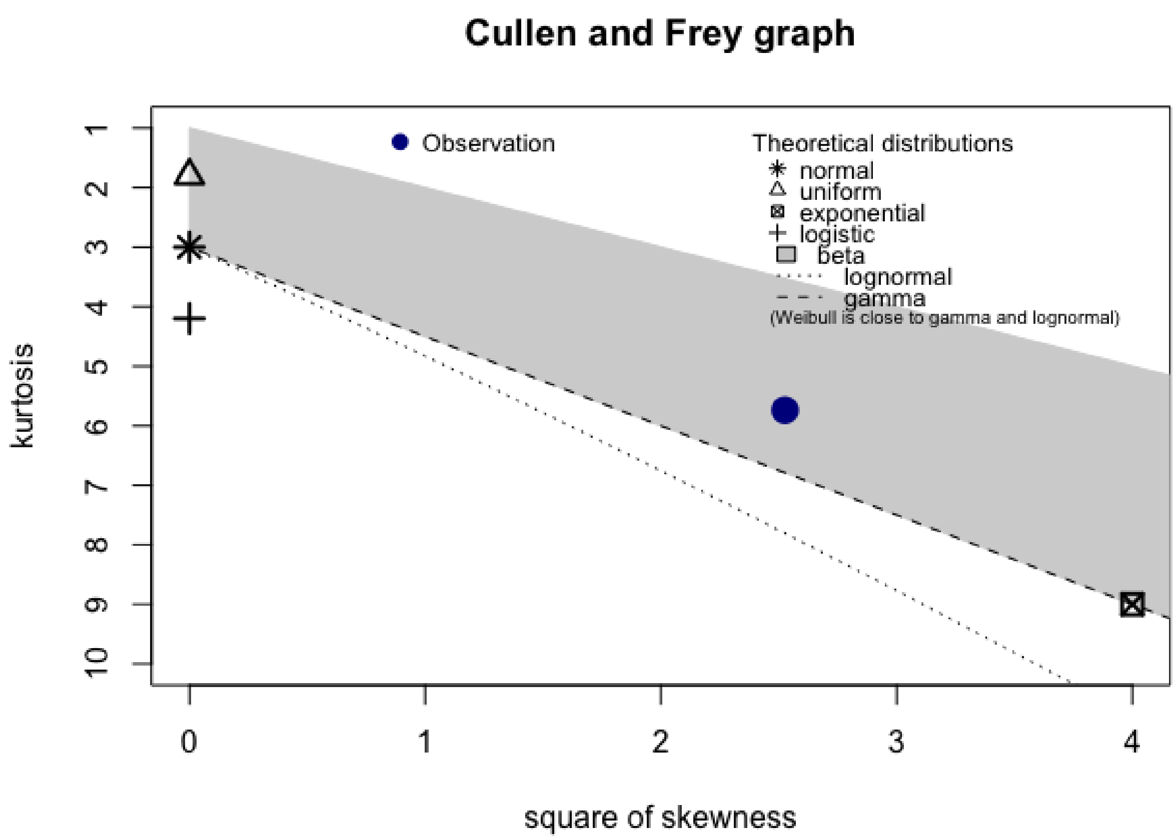
mean: 219.3074

estimated sd: 100.7125

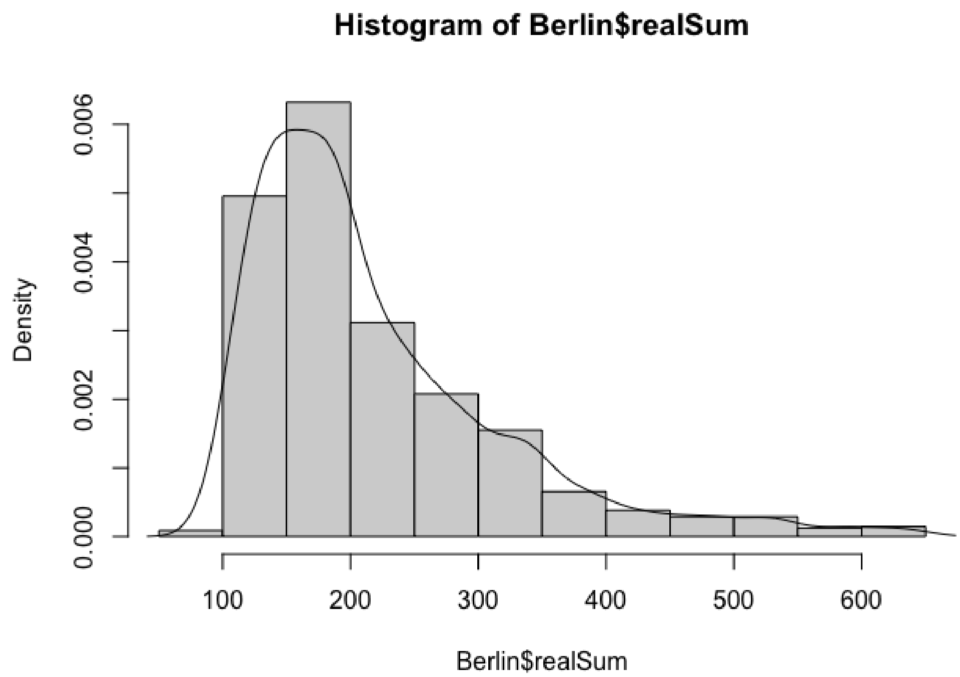
estimated skewness: 1.589429

estimated kurtosis: 5.738424

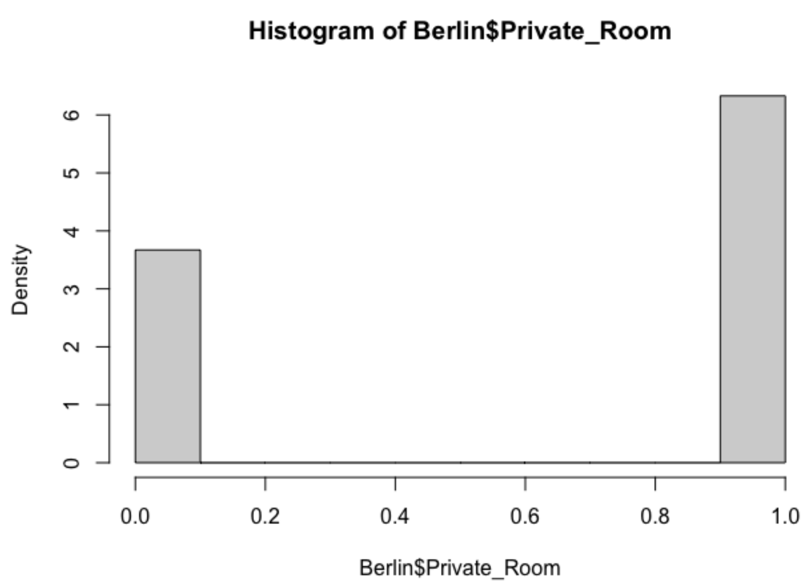
SM1



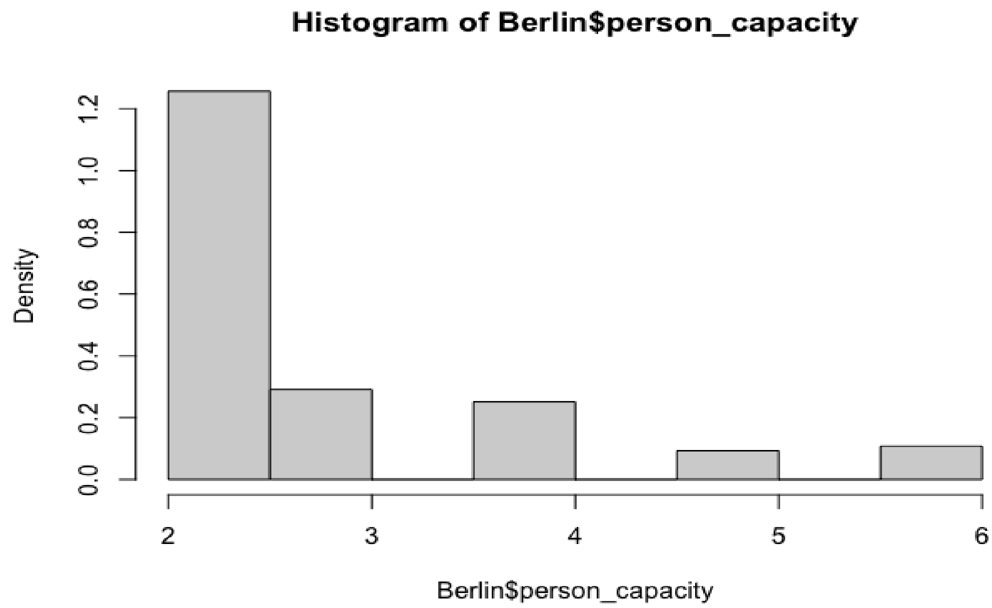
CF1


Histogram of Berlin$realSum
Density
0.000
0.002
0.004
0.006
100
200
300
400
500
600
Berlin$realSum

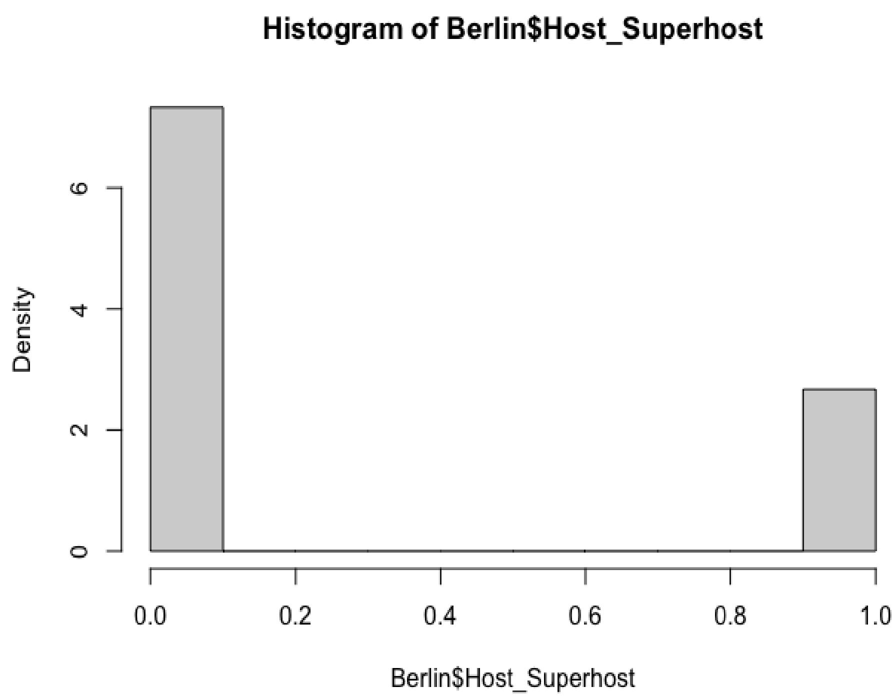
HT1


Histogram of Berlin$Private_Room
Density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
Berlin$Private_Room

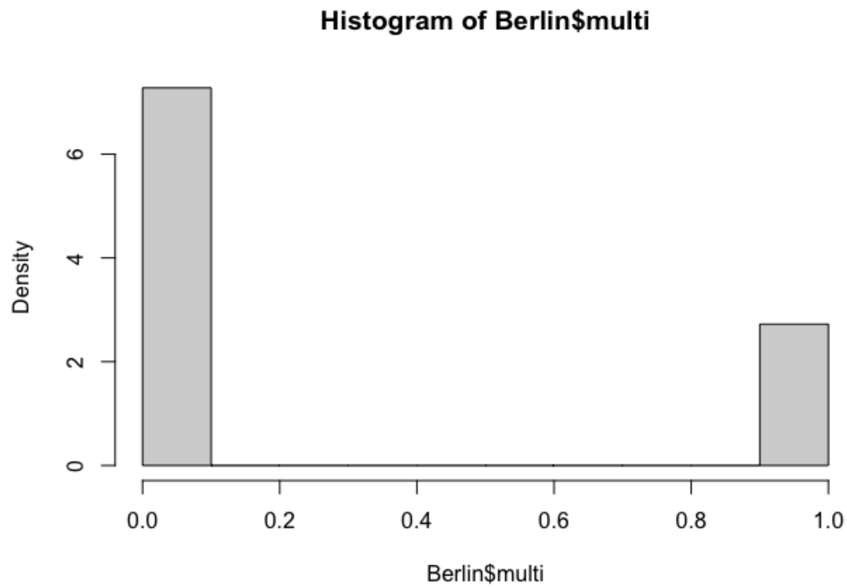
HT2


Histogram of Berlin$person_capacity
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2
3
4
5
6
Berlin$person_capacity

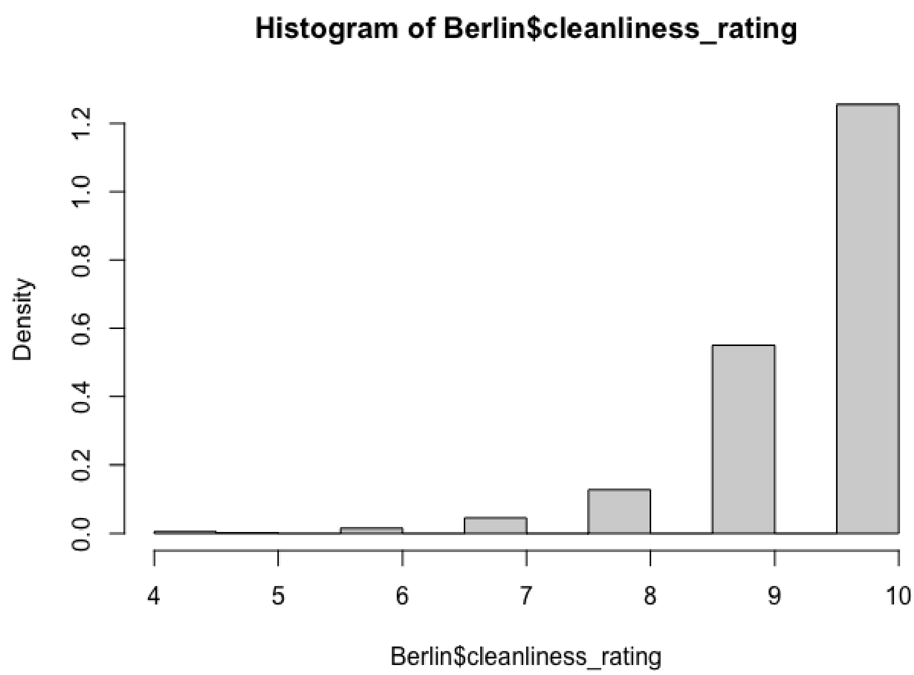
HT3


Histogram of Berlin$Host_Superhost
Density
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
Berlin$Host_Superhost

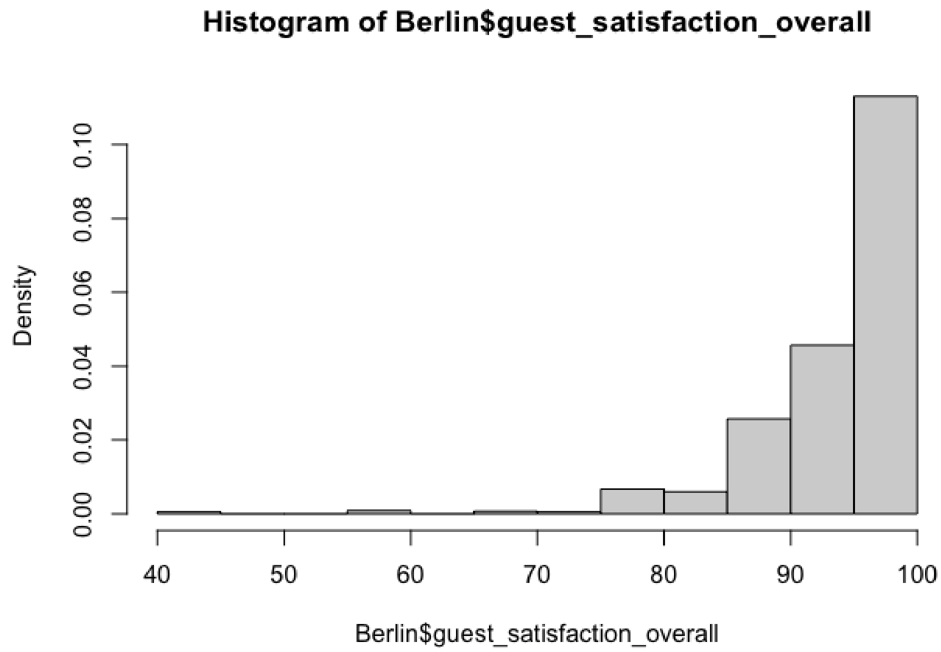
HT4


Histogram of Berlin$multi
Density
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
Berlin$multi

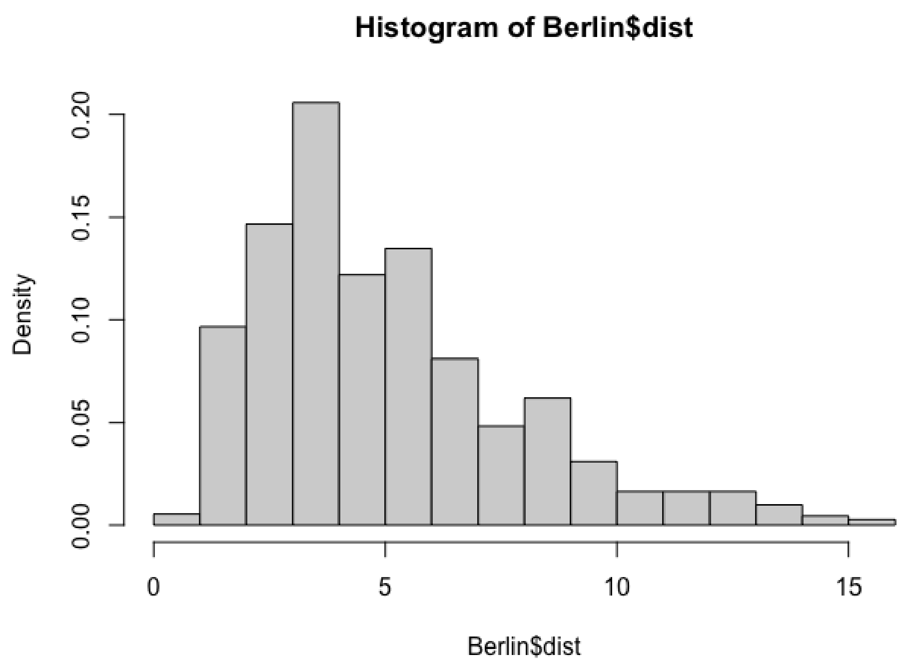
HT5


Histogram of Berlin$cleanliness_rating
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
4
5
6
7
8
9
10
Berlin$cleanliness_rating

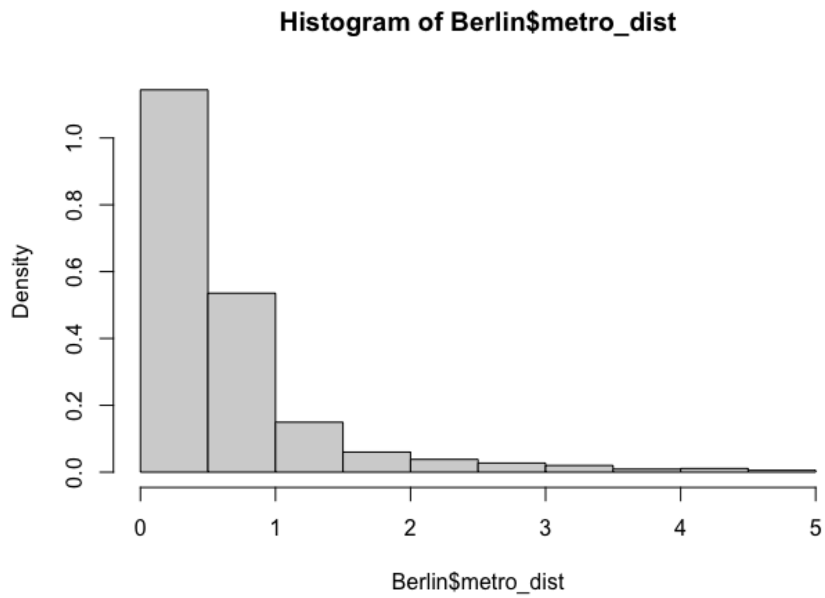
HT6


Histogram of Berlin$guest_satisfaction_overall
Density
0.00
0.02
0.04
0.06
0.08
0.10
40
50
60
70
80
90
100
Berlin$guest_satisfaction_overall

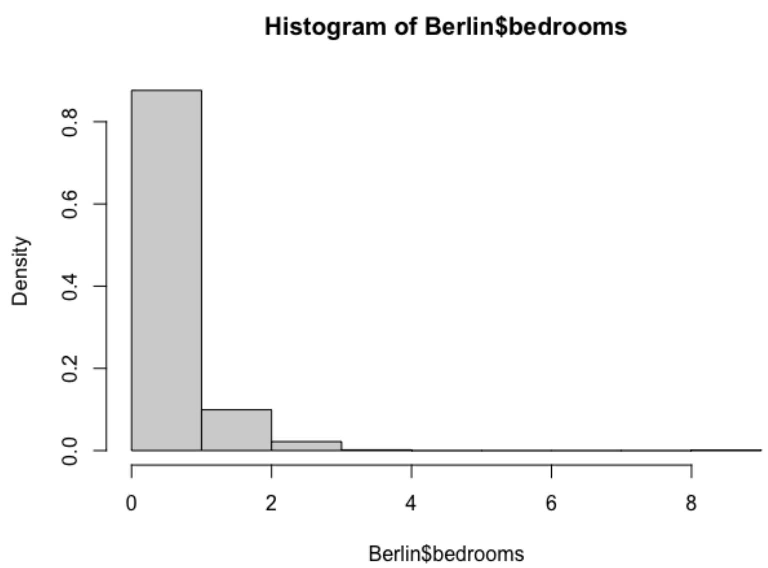
HT7


Histogram of Berlin$dist
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
Berlin$dist

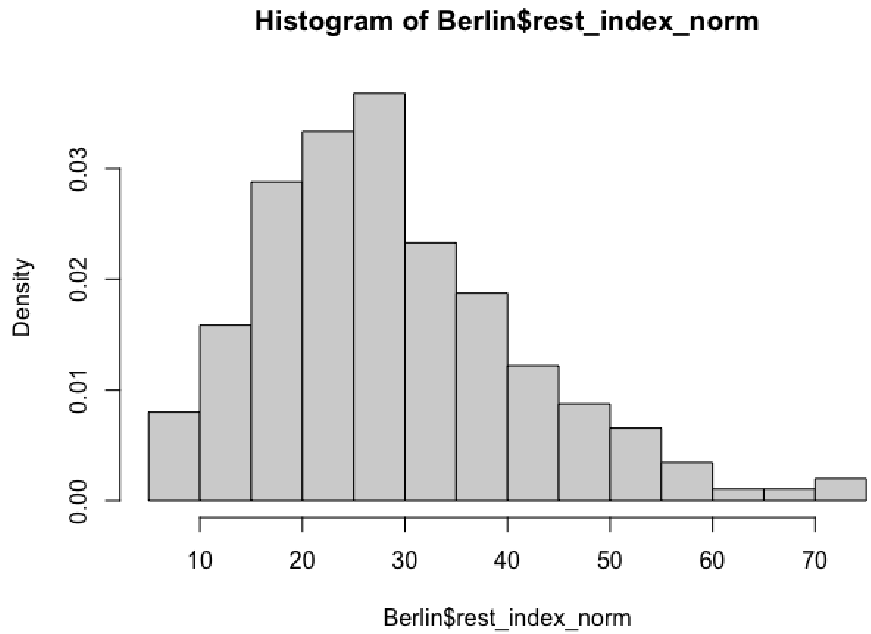
HT8


Histogram of Berlin$metro_dist
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
Berlin$metro_dist

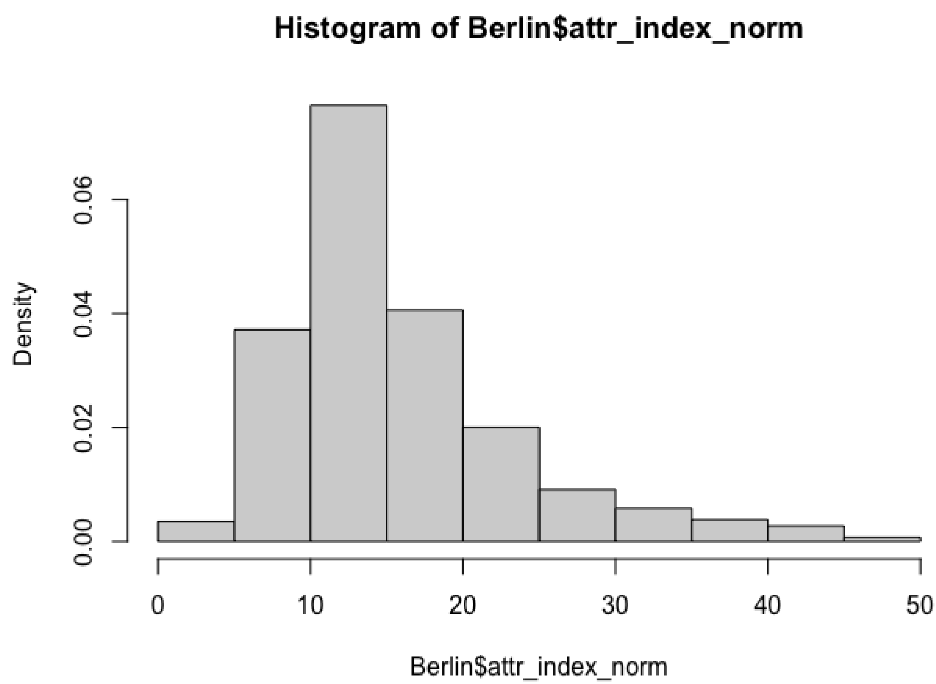
HT9


Histogram of Berlin$bedrooms
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
Berlin$bedrooms

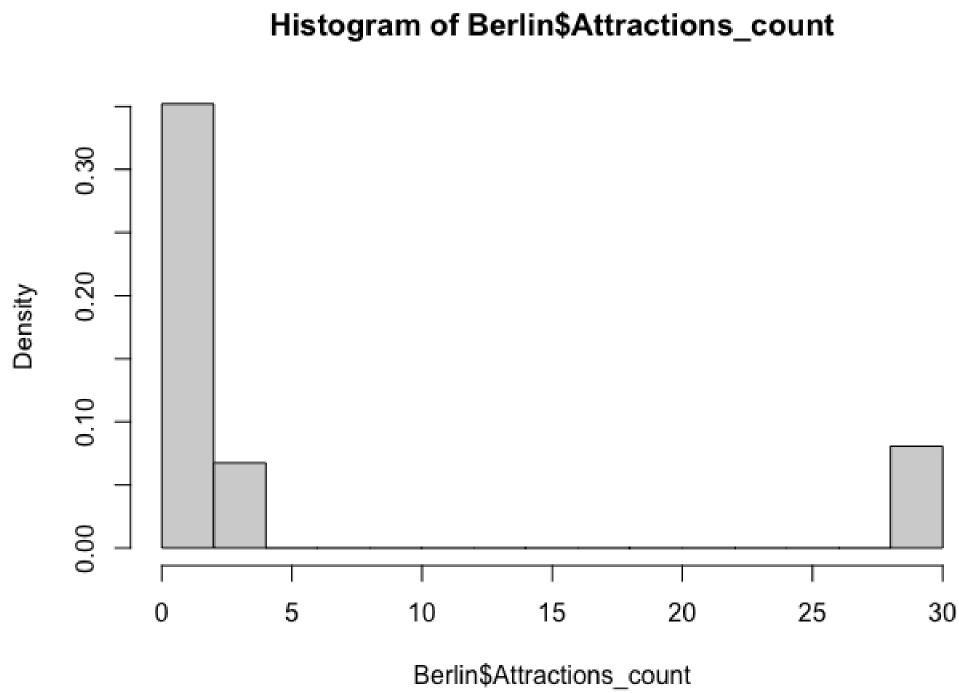
HT10


Histogram of Berlin$rest_index_norm
Density
0.00
0.01
0.02
0.03
10
20
30
40
50
60
70
Berlin$rest_index_norm

HT11


Histogram of Berlin$attr_index_norm
Density
0.00
0.02
0.04
0.06
0
10
20
30
40
50
Berlin$attr_index_norm

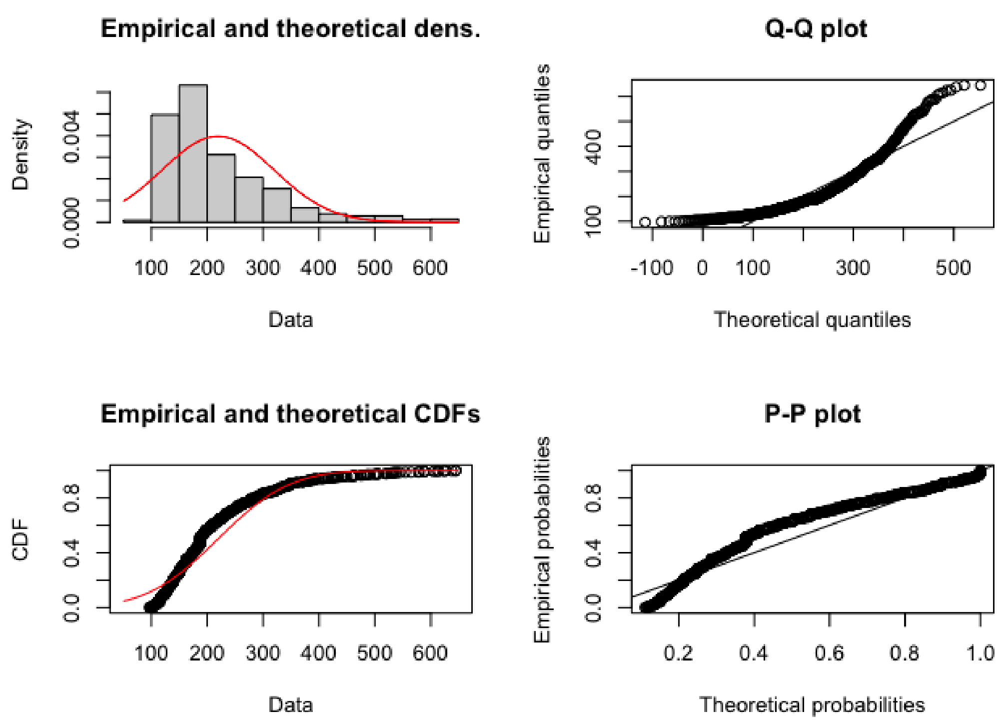
HT12


Histogram of Berlin$Attractions_count
Density
0.00
0.10
0.20
0.30
0
5
10
15
20
25
30
Berlin$Attractions_count

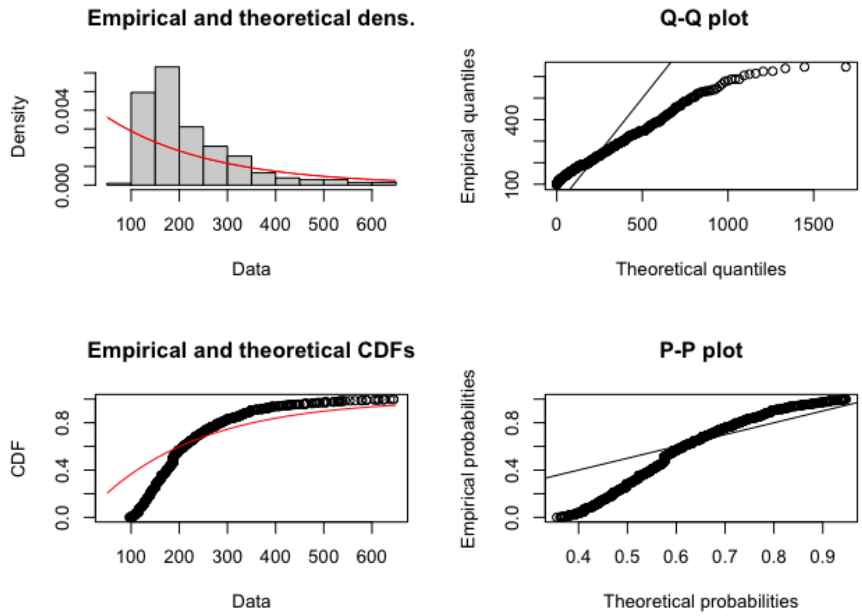
HT13

	A	B	C	D	E	F
1	realSum	Private_Room	person_capac	Host_Superhc	multi	biz
2	644.806955	0	6	1	1	
3	644.806955	0	6	1	1	
4	637.795644	0	4	0	1	
5	624.941572	0	6	0	0	
6	624.006731	0	5	0	0	
7	614.424605	0	5	1	1	
8	611.15266	0	5	0	0	
9	605.3099	0	2	0	1	
10	588.015331	0	6	0	1	
11	587.781621	0	5	0	0	
12	587.781621	0	5	0	0	
13	579.835468	0	6	0	0	
14	577.498364	0	3	0	0	
15	564.878003	0	5	0	0	
16	554.361036	0	4	0	0	
17	545.01262	0	3	1	1	
18	539.169861	0	2	1	0	
19	539.169861	0	5	1	1	
20	534.495653	0	5	0	0	
21	533.327101	0	2	0	0	
22	532.39226	0	6	0	1	

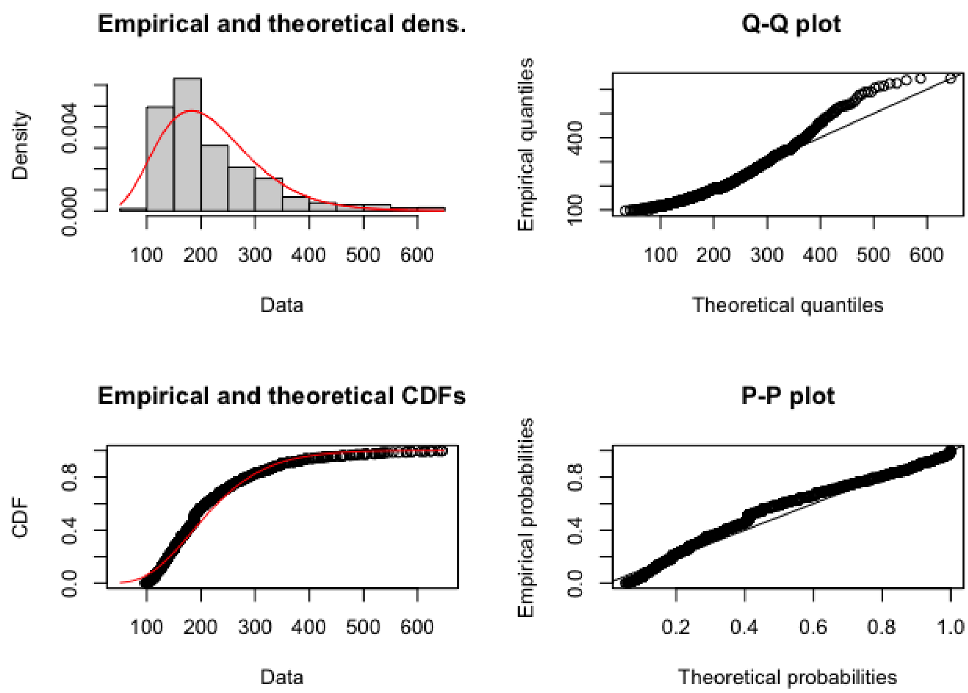
CND



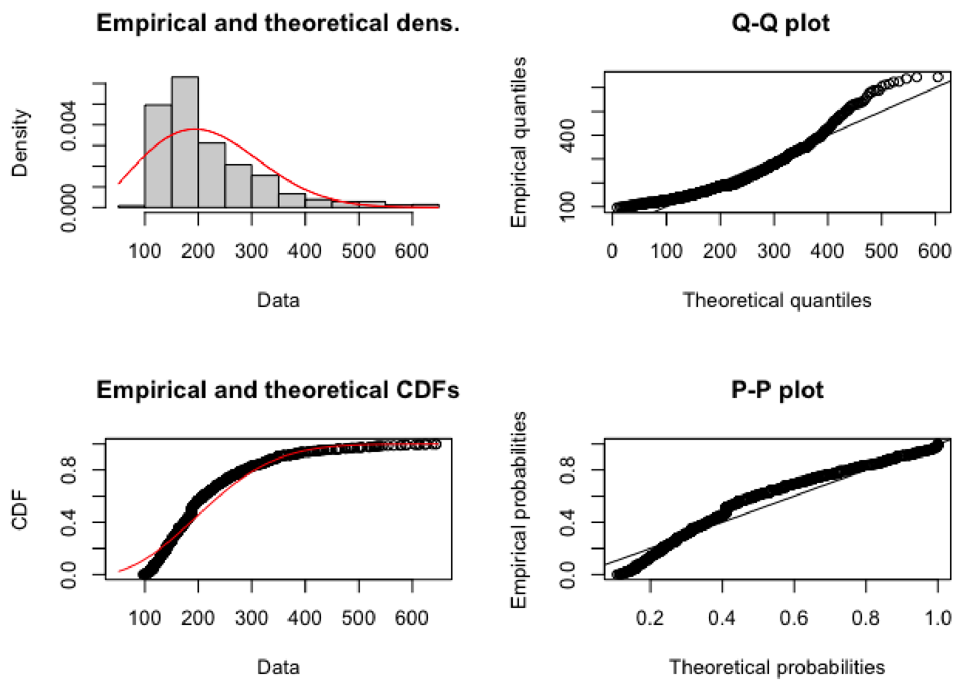
DP1


Empirical and theoretical dens.
Density
Data
Q-Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P-P plot
Empirical probabilities
Theoretical probabilities

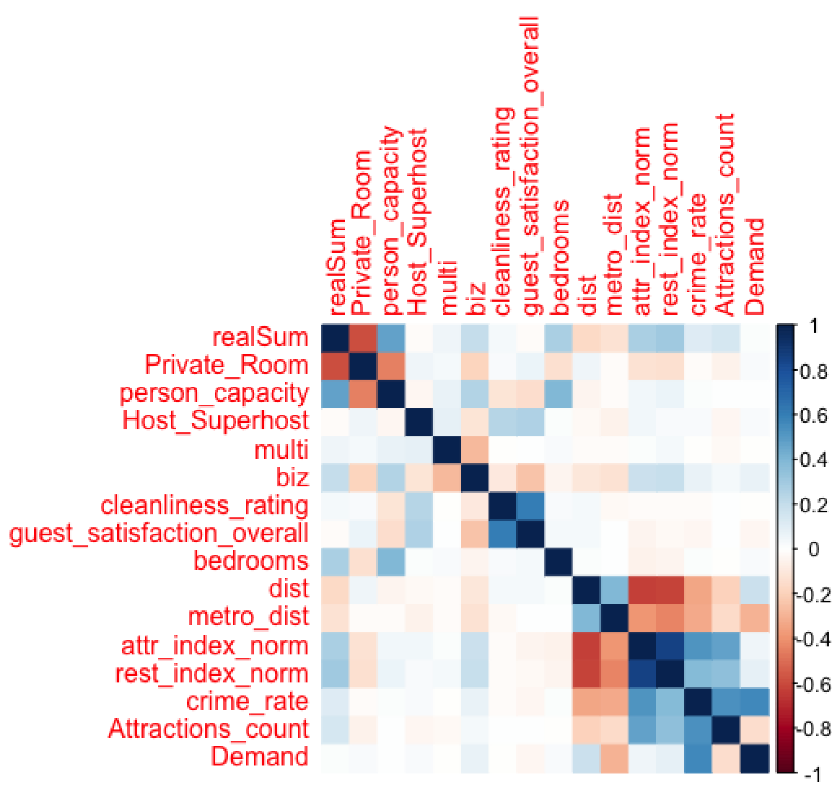
DP2


Empirical and theoretical dens.
Density
Data
Q-Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P-P plot
Empirical probabilities
Theoretical probabilities

DP3


Empirical and theoretical dens.
Density
Data
Q-Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P-P plot
Empirical probabilities
Theoretical probabilities

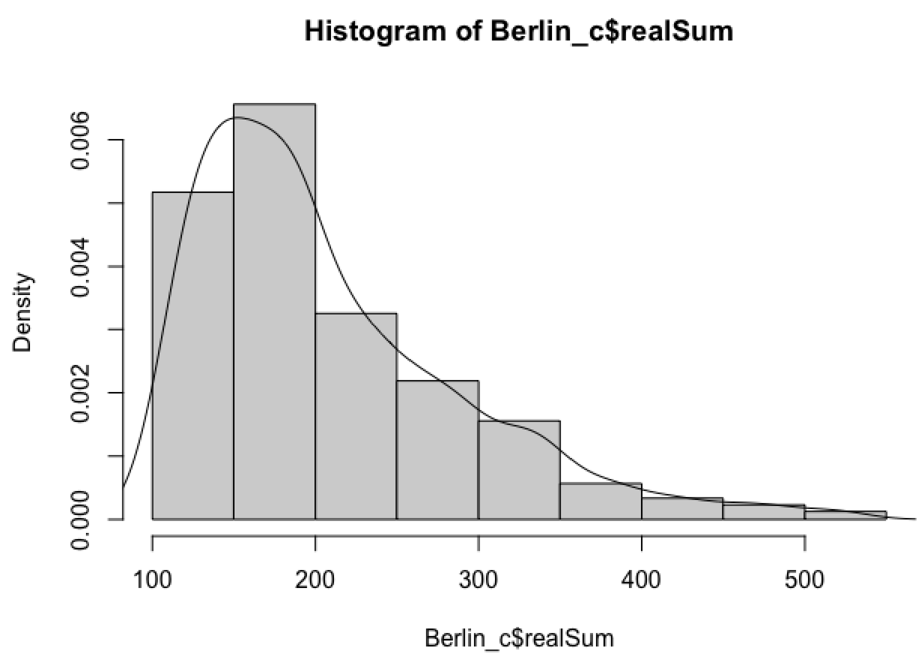
DP4


realSum
Private_Room
person_capacity
Host_Superhost
multi
biz
cleanliness_rating
guest_satisfaction_overall
bedrooms
dist
metro_dist
attr_index_norm
rest_index_norm
crime_rate
Attractions_count
Demand

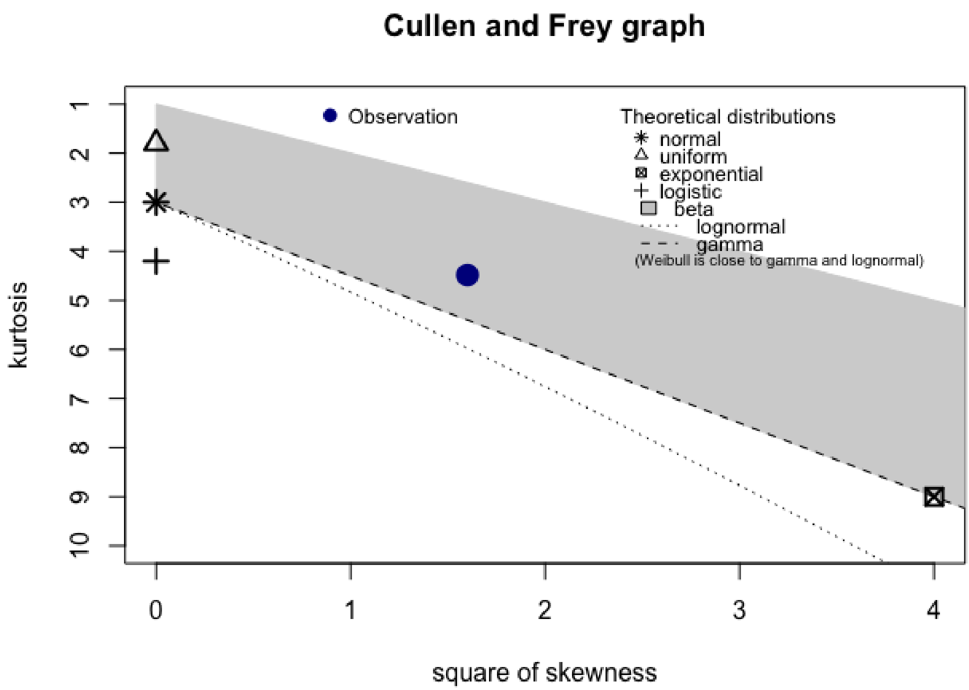
CM1

```
> descdist(Berlin_c$realSum, discrete = FALSE)
summary statistics
------
min:  100.7292   max:  530.99
median:  187.7863
mean:  208.5591
estimated sd:  82.80326
estimated skewness:  1.265235
estimated kurtosis:  4.482459
```

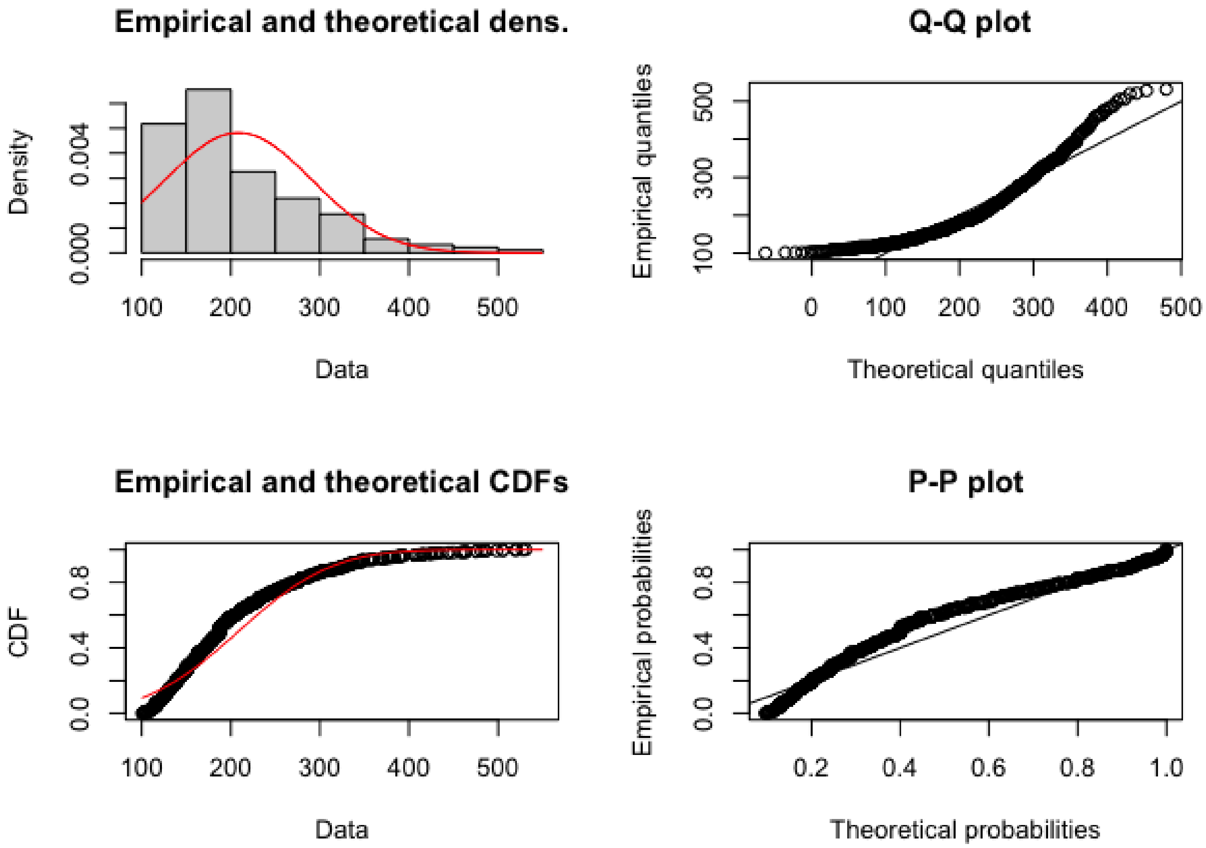
SM2



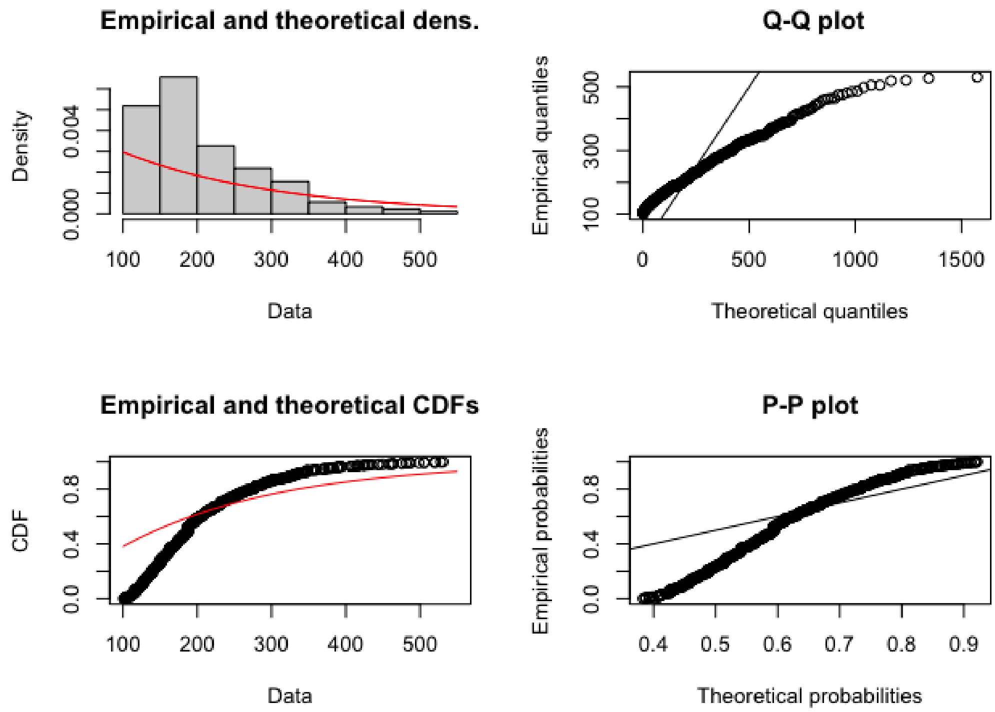
YHT


Cullen and Frey graph
Observation
Theoretical distributions
normal
uniform
exponential
logistic
beta
lognormal
gamma
(Weibull is close to gamma and lognormal)
kurtosis
square of skewness
1
2
3
4
5
6
7
8
9
10
0
1
2
3
4

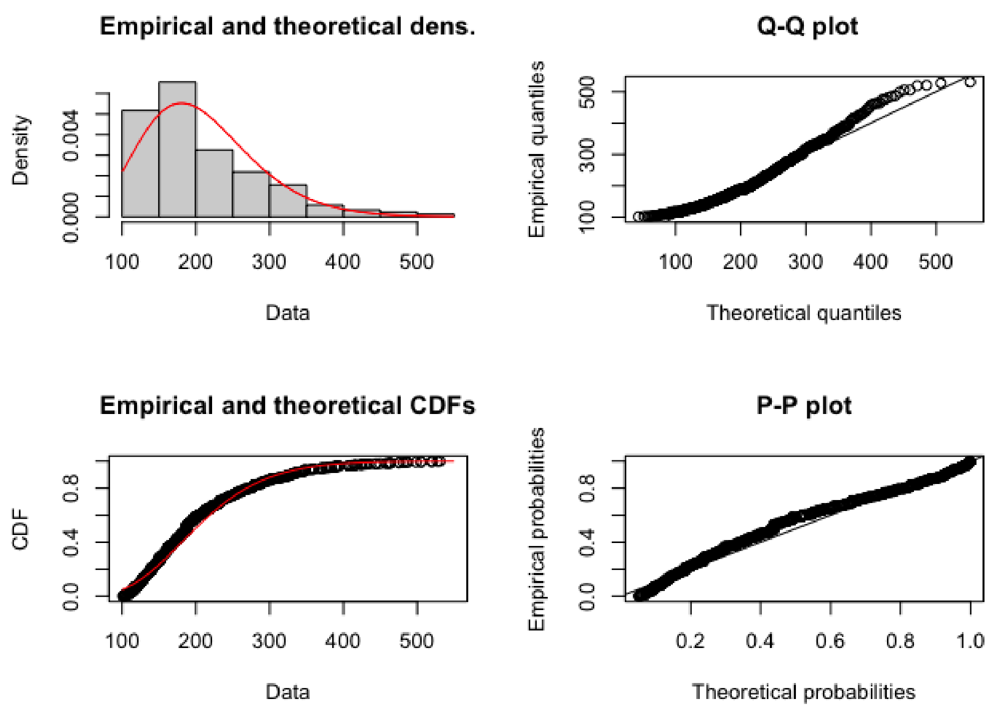
CF2


Empirical and theoretical dens.
Density
0.000
0.004
100
200
300
400
500
Data
Q-Q plot
Empirical quantiles
100
300
500
0
100
200
300
400
500
Theoretical quantiles
Empirical and theoretical CDFs
CDF
0.0
0.4
0.8
100
200
300
400
500
Data
P-P plot
Empirical probabilities
0.0
0.4
0.8
0.2
0.4
0.6
0.8
1.0
Theoretical probabilities

DP5


Empirical and theoretical dens.
Density
Data
Q-Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P-P plot
Empirical probabilities
Theoretical probabilities

DP6


Empirical and theoretical dens.
Density
Data
Q-Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P-P plot
Empirical probabilities
Theoretical probabilities

DP7

```
summary(Berlin_c$cleanliness_rating)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 8.000   9.000  10.000   9.601  10.000  10.000

summary(Berlin_c$guest_satisfaction_overall)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 80.00   93.00   97.00   95.32   99.00  100.00

summary(Berlin_c$bedrooms)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   1.000   1.000   1.024   1.000   2.000

summary(Berlin_c$dist)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.011   3.077   4.323   4.925   6.139  13.846
```

NA1

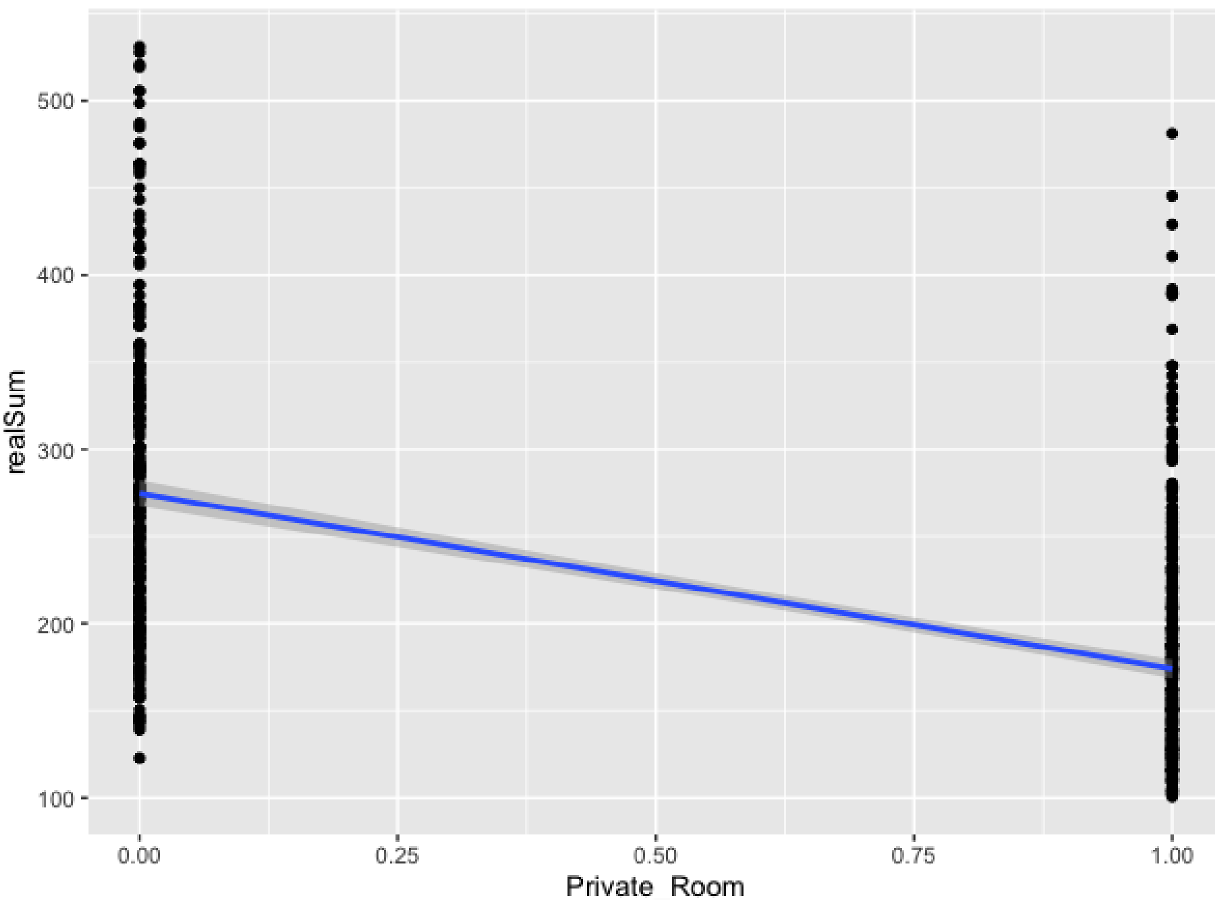
```
summary(Berlin_c$metro_dist)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.09929 0.29823 0.45165 0.65835 0.74406 4.63648

summary(Berlin_c$attr_index_norm)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 5.012  10.829  13.707  15.048  17.980  38.471

summary(Berlin_c$rest_index_norm)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 8.811  19.819  26.476  27.991  34.789  59.849

Crime rate
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 7.494   8.914  12.562  13.074  18.046  18.795

summary(Berlin_c$Attractions_count)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   0.000   0.000   4.905   3.000  30.000

summary(Berlin_c$Demand)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.01875  0.00000  0.07500  0.05992  0.11250  0.11250
```

NA2

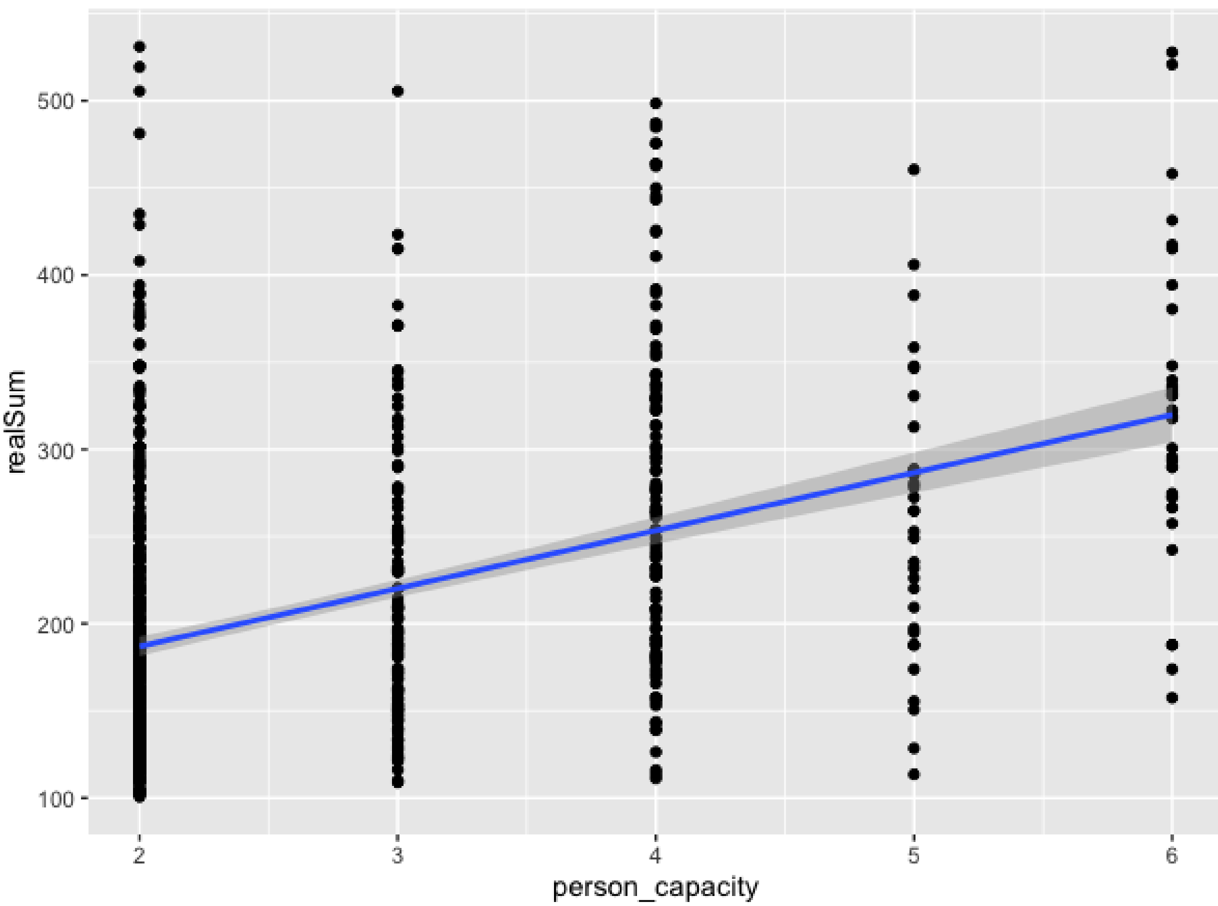
Dummy and Categorical Variables Analysis

```
tabulate(Berlin_c$Private_Room)
[1] 626

tabulate(Berlin_c$person_capacity)
[1]   0 623 143 115  35  35

tabulate(Berlin_c$Host_Superhost)
[1] 261

tabulate(Berlin_c$multi)
[1] 250

tabulate(Berlin_c$biz)
[1] 160
```

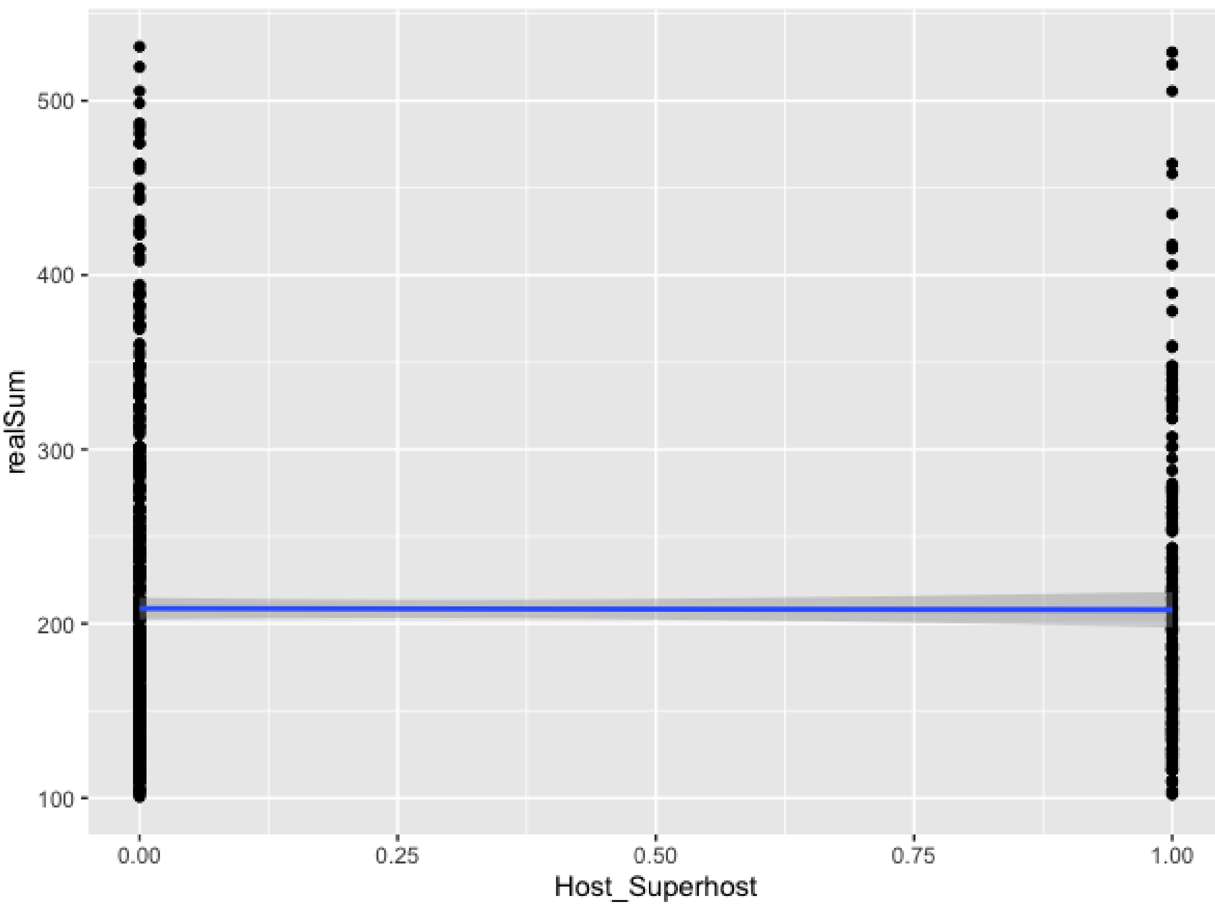
DV



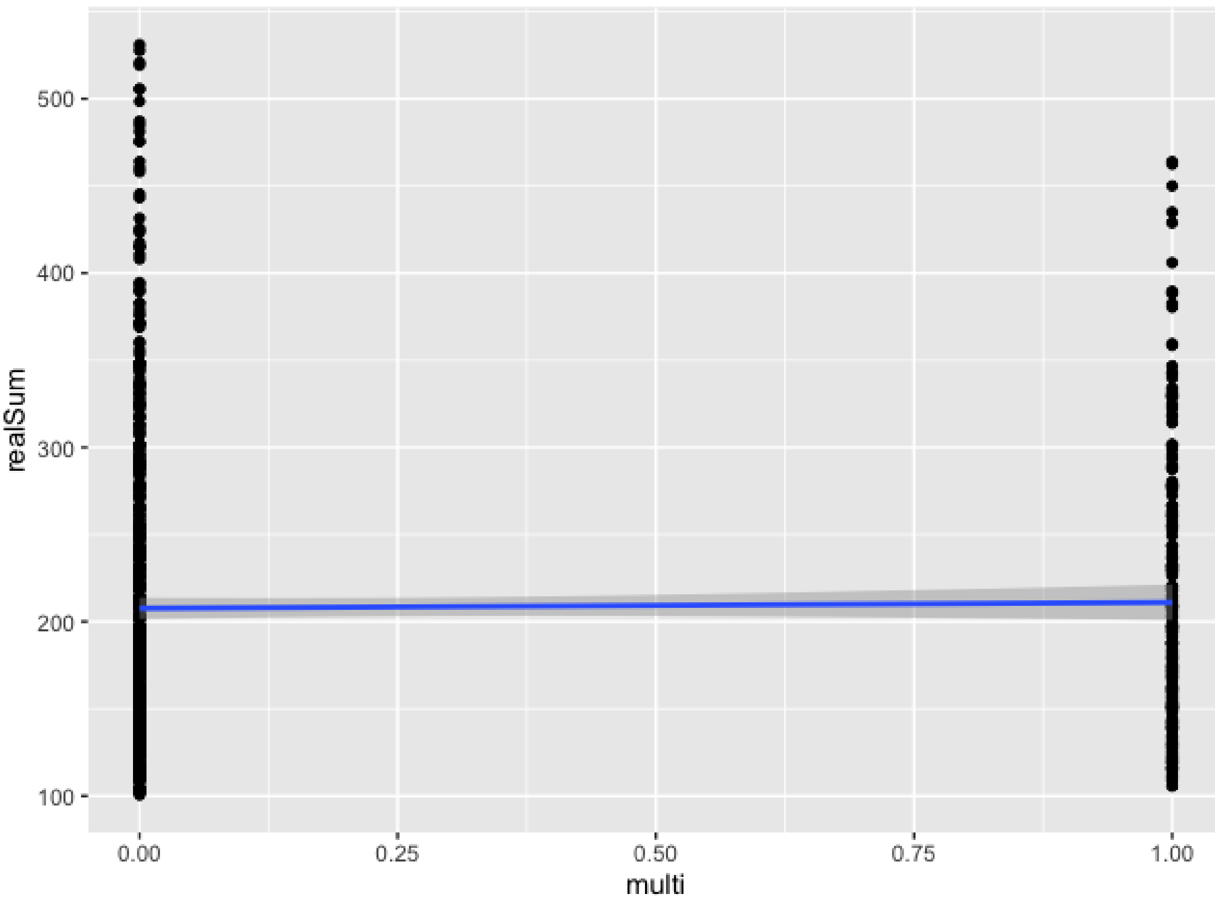
BV1



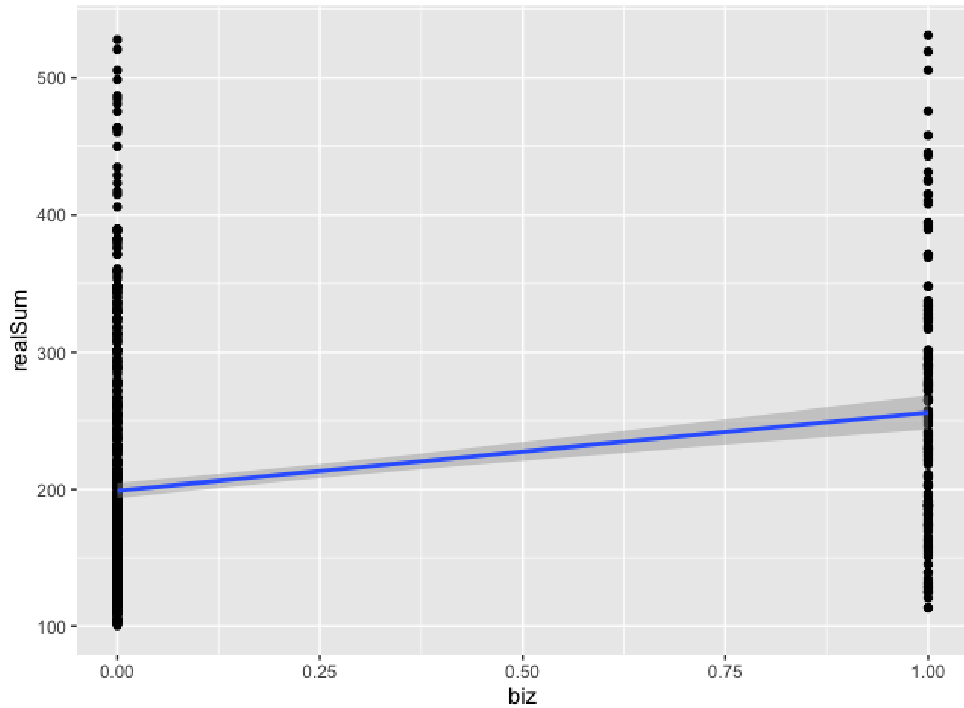
BV2


500
400
300
200
100
realSum
0.00
0.25
0.50
0.75
1.00
Host_Superhost

BV3

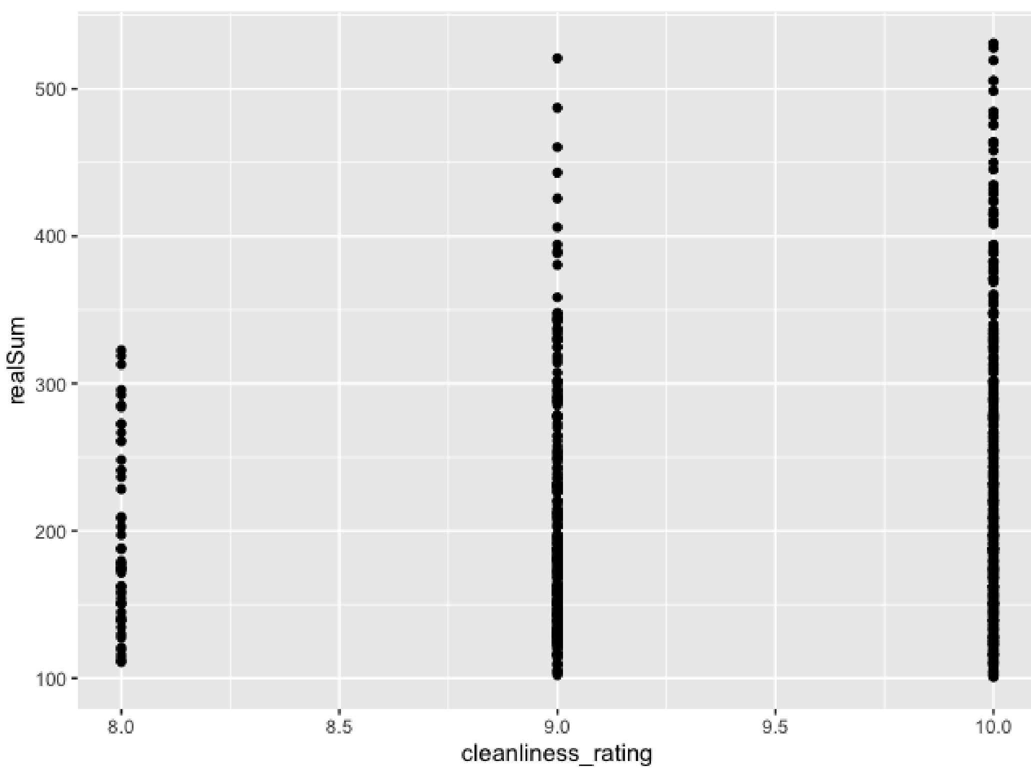


BV4



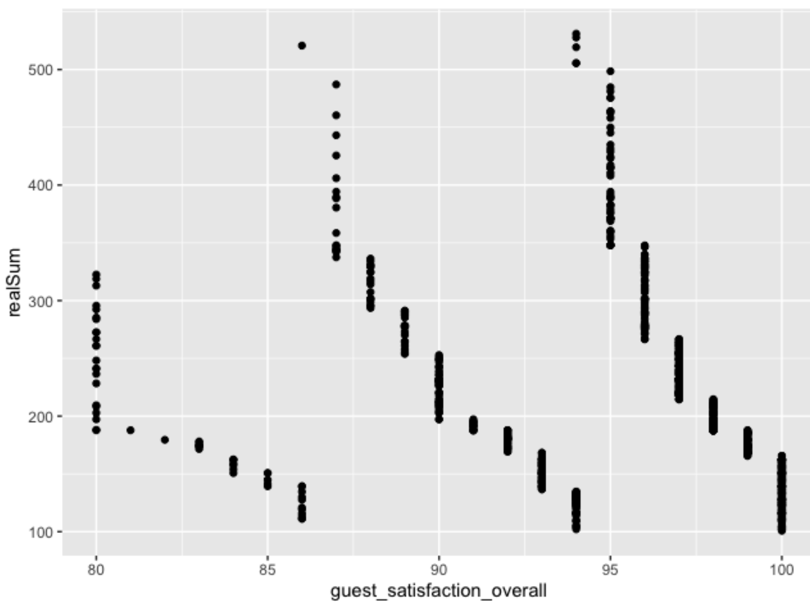
realSum
500
400
300
200
100
0.00
0.25
0.50
0.75
1.00
biz

BV5



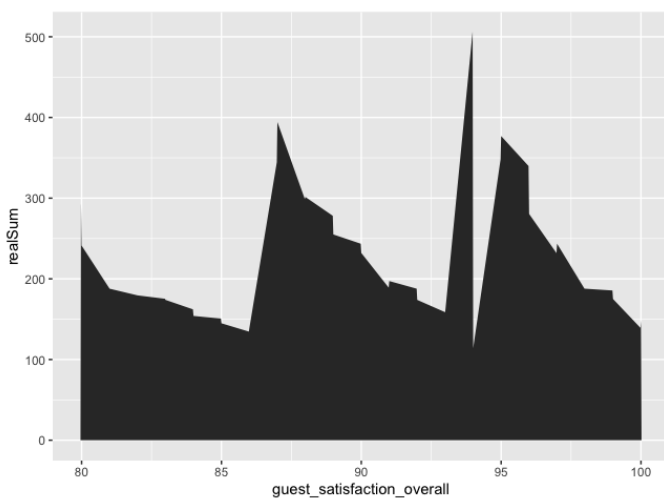
realSum
500
400
300
200
100
8.0
8.5
9.0
9.5
10.0
cleanliness_rating

BV6



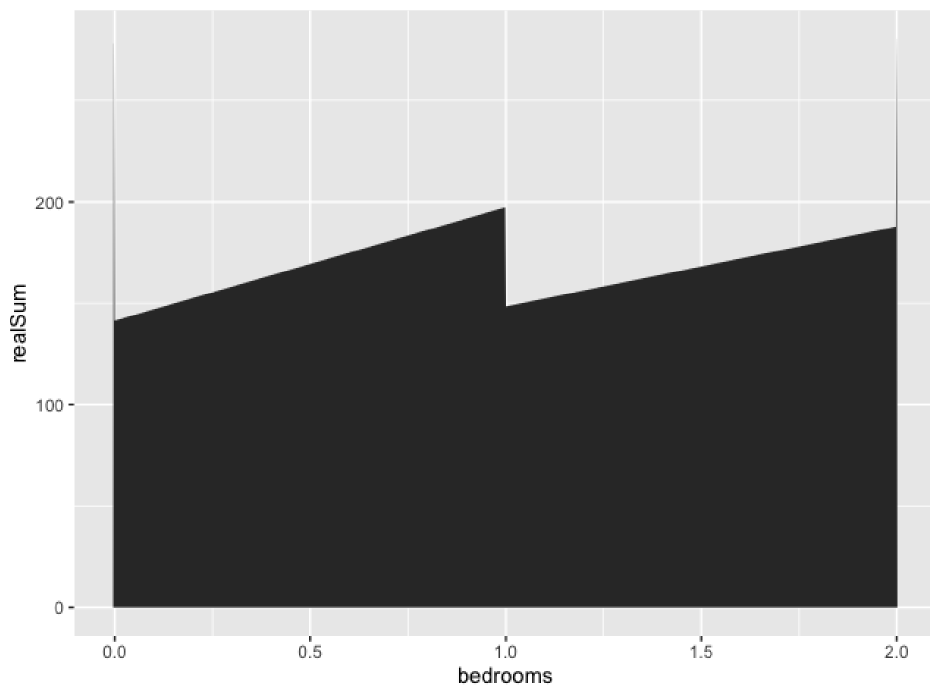
500
400
300
200
100
realSum
80
85
90
95
100
guest_satisfaction_overall

BV7

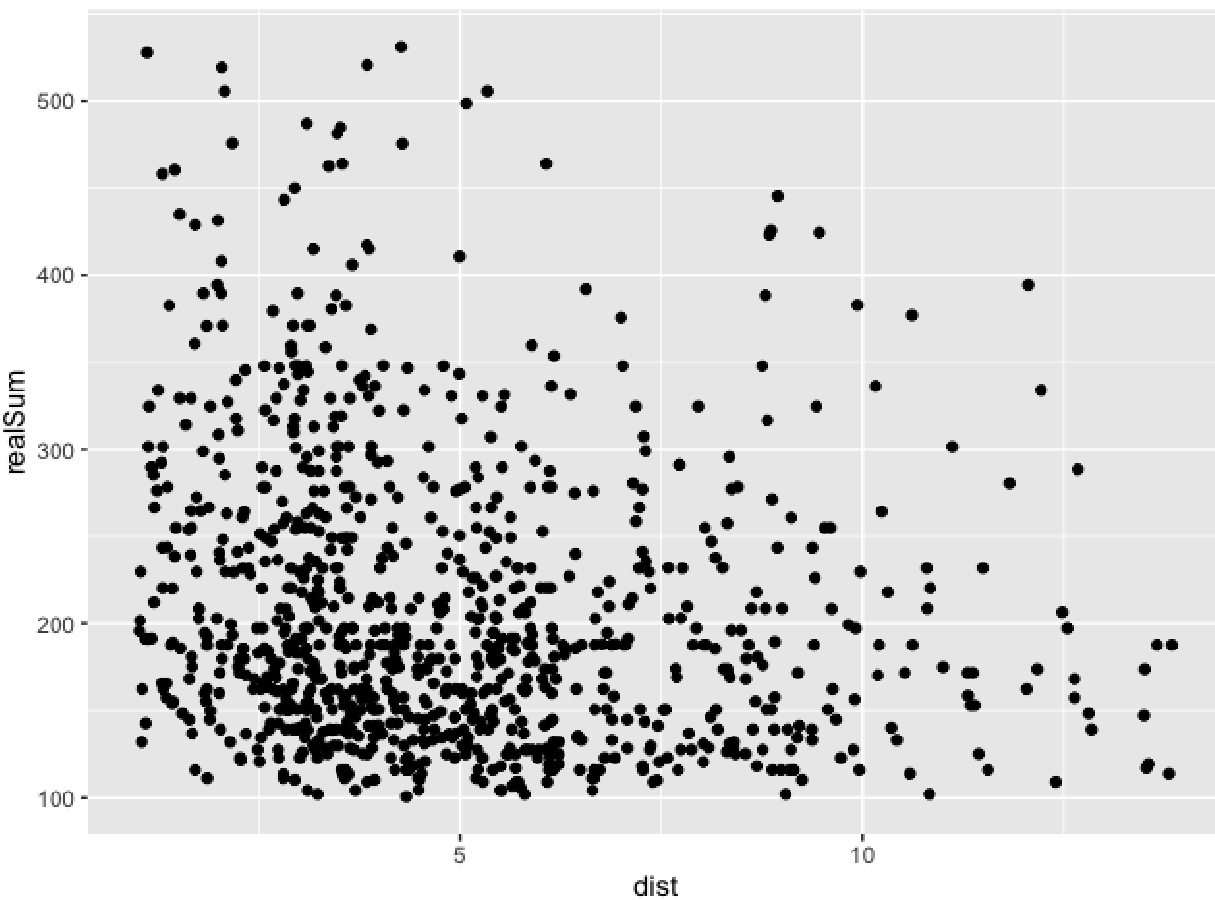


500
400
300
200
100
0
realSum
80
85
90
95
100
guest_satisfaction_overall

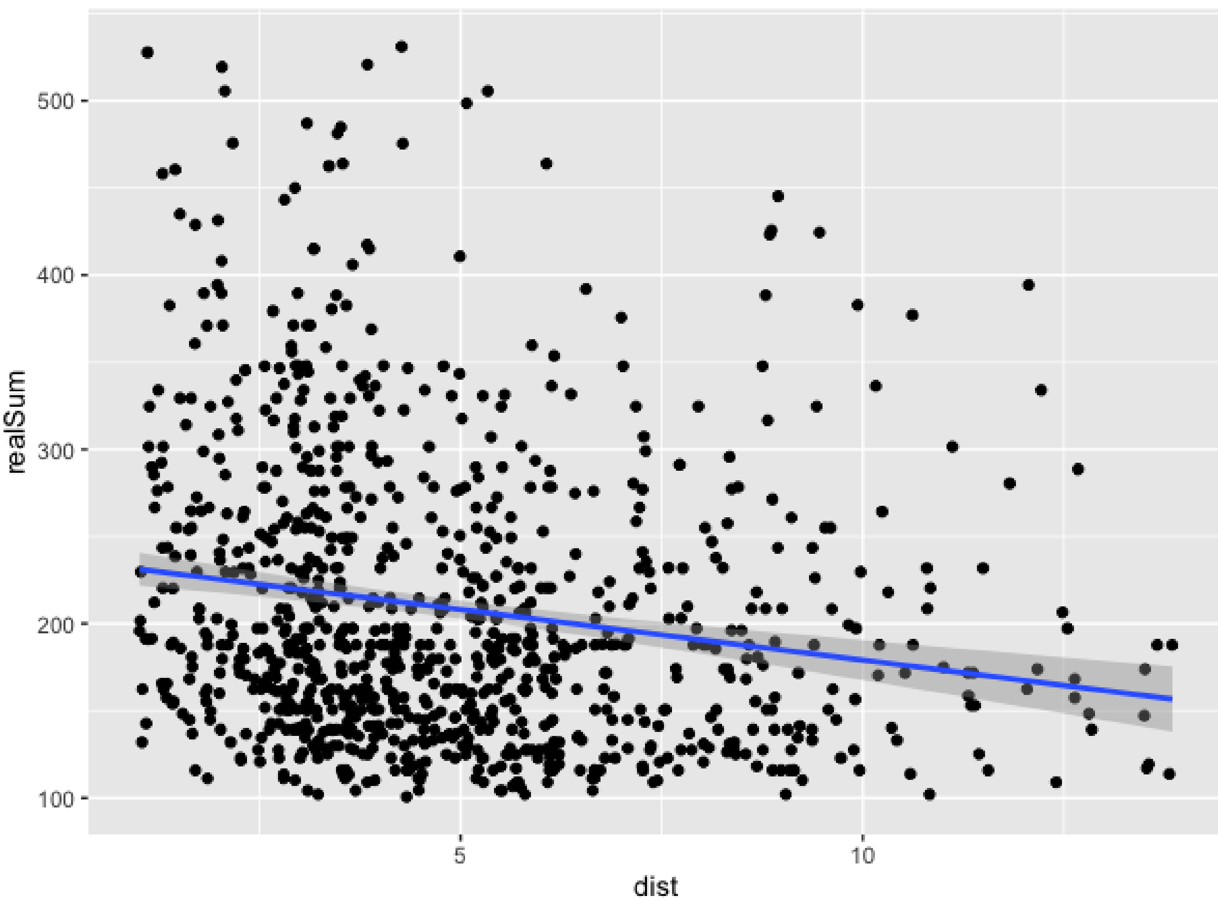
BV8



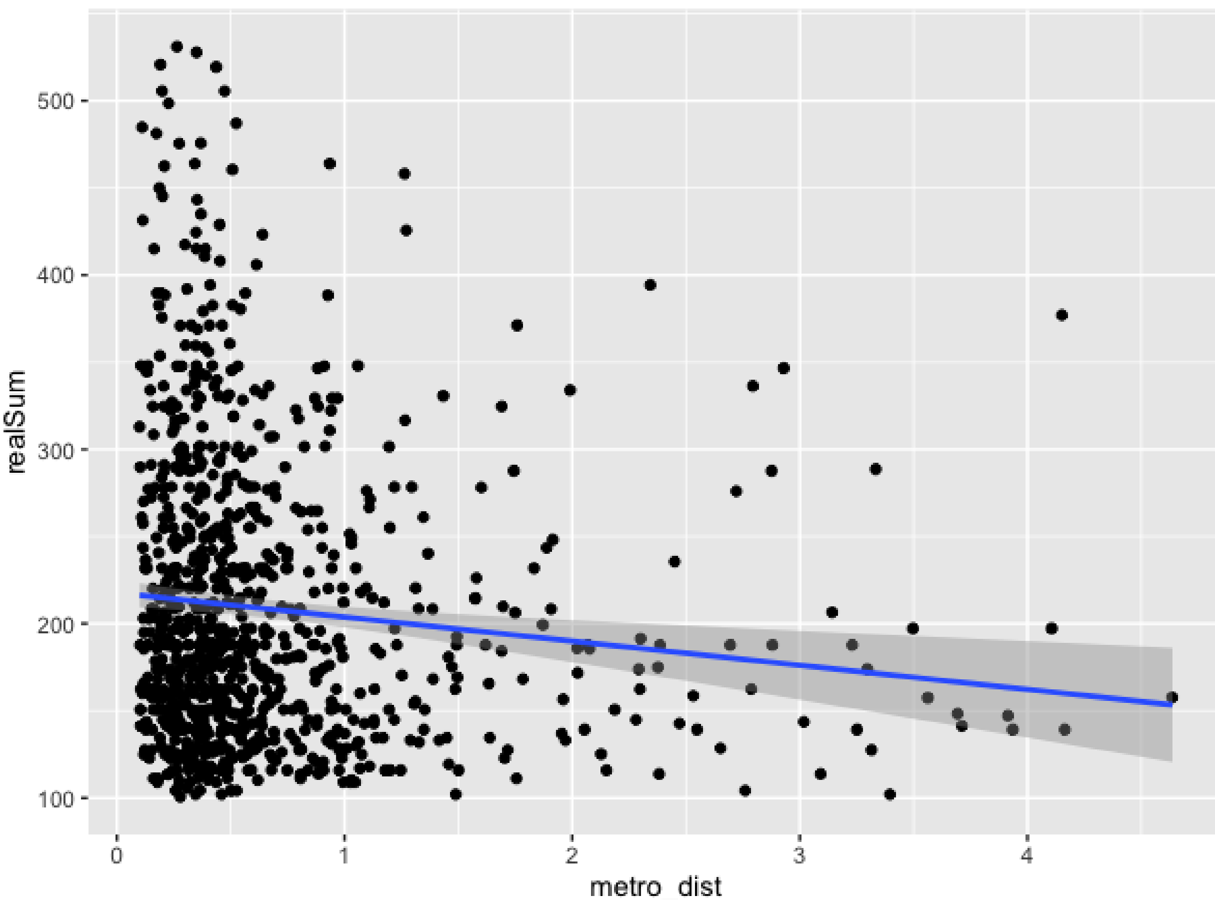
BV9



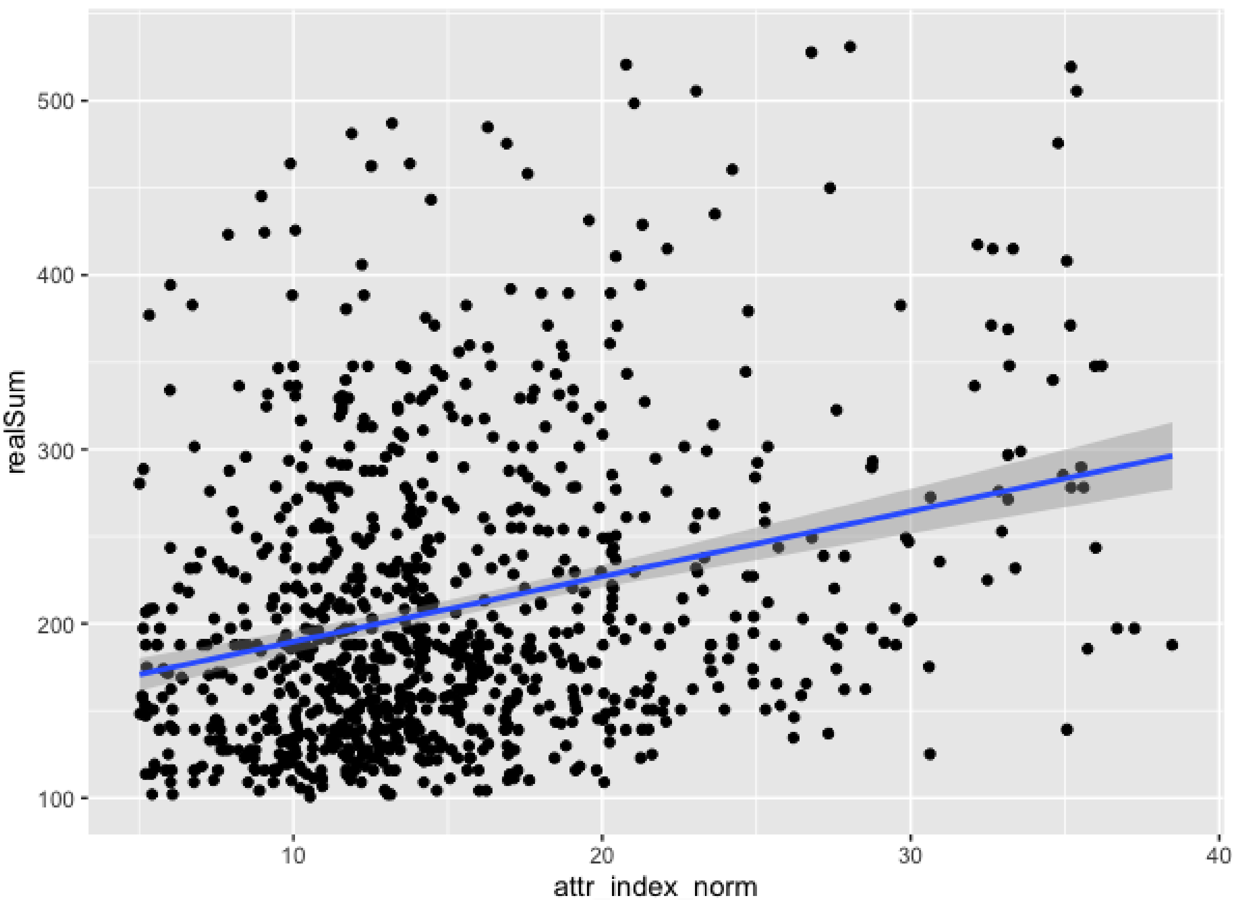
BV10



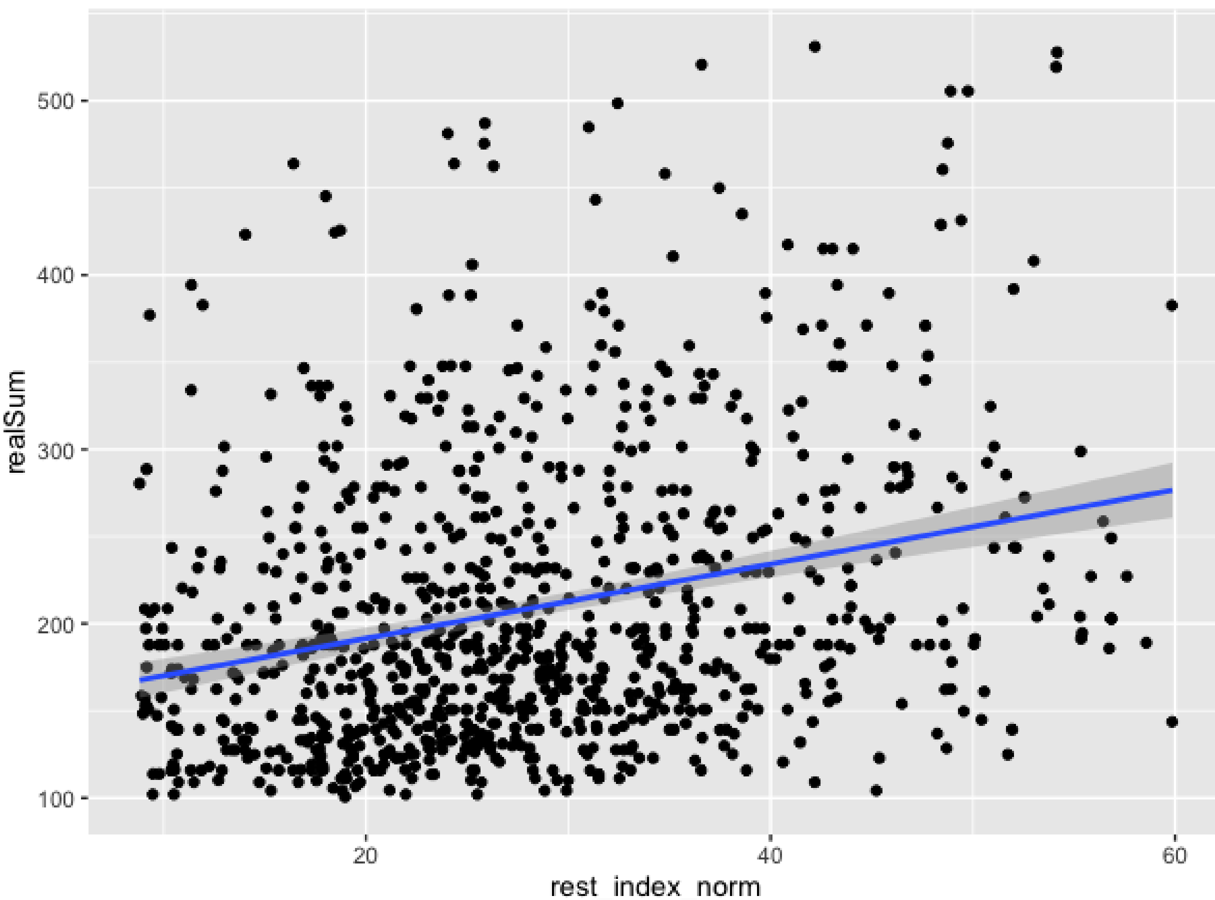
BV11



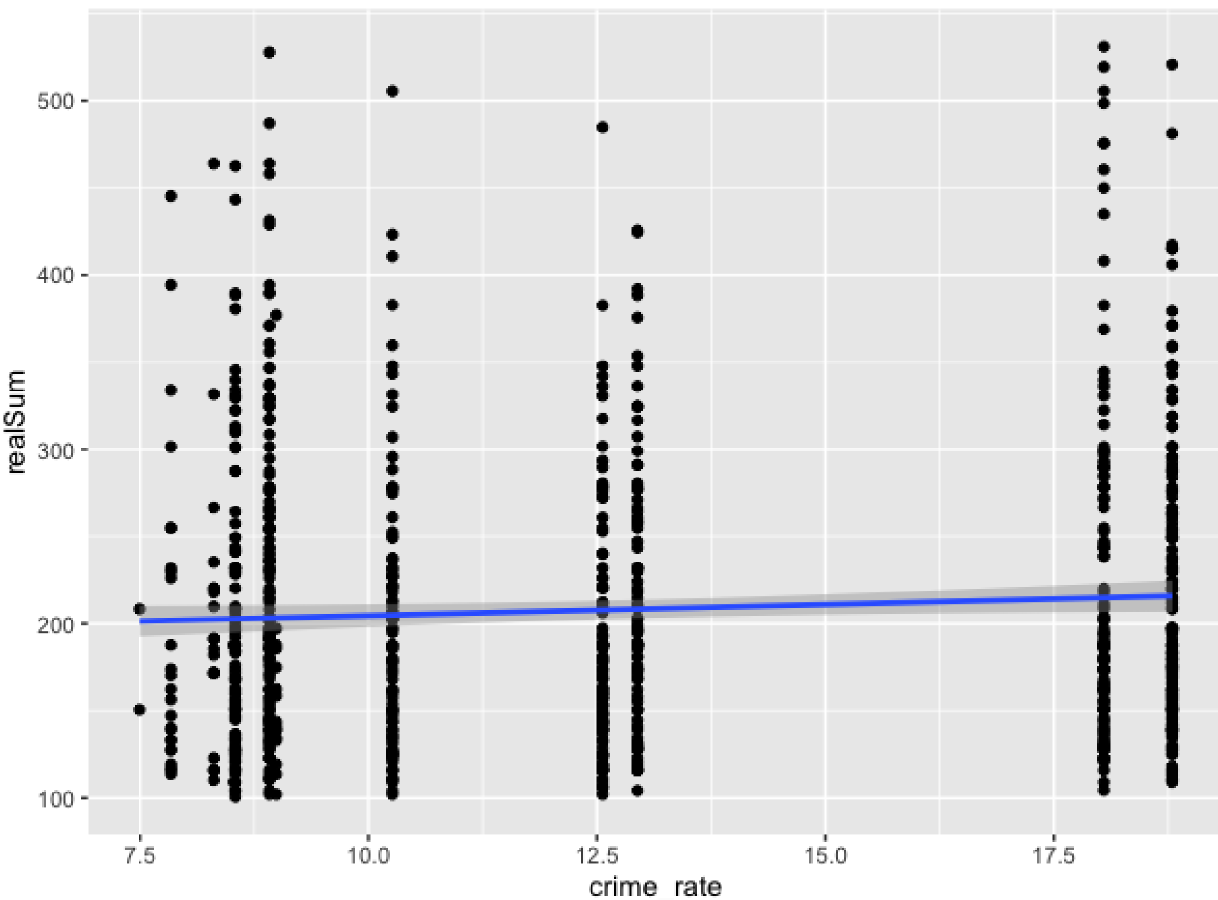
BV12



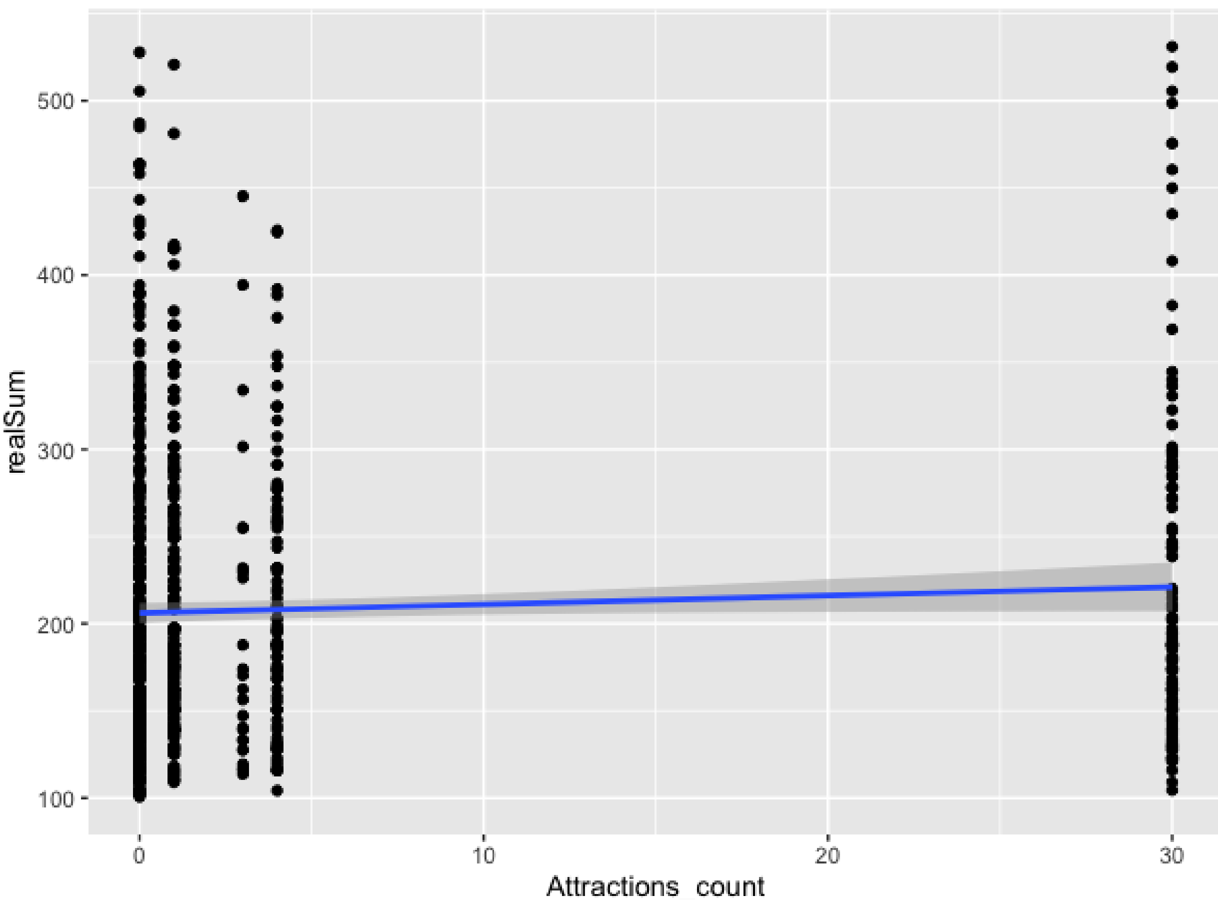
BV13



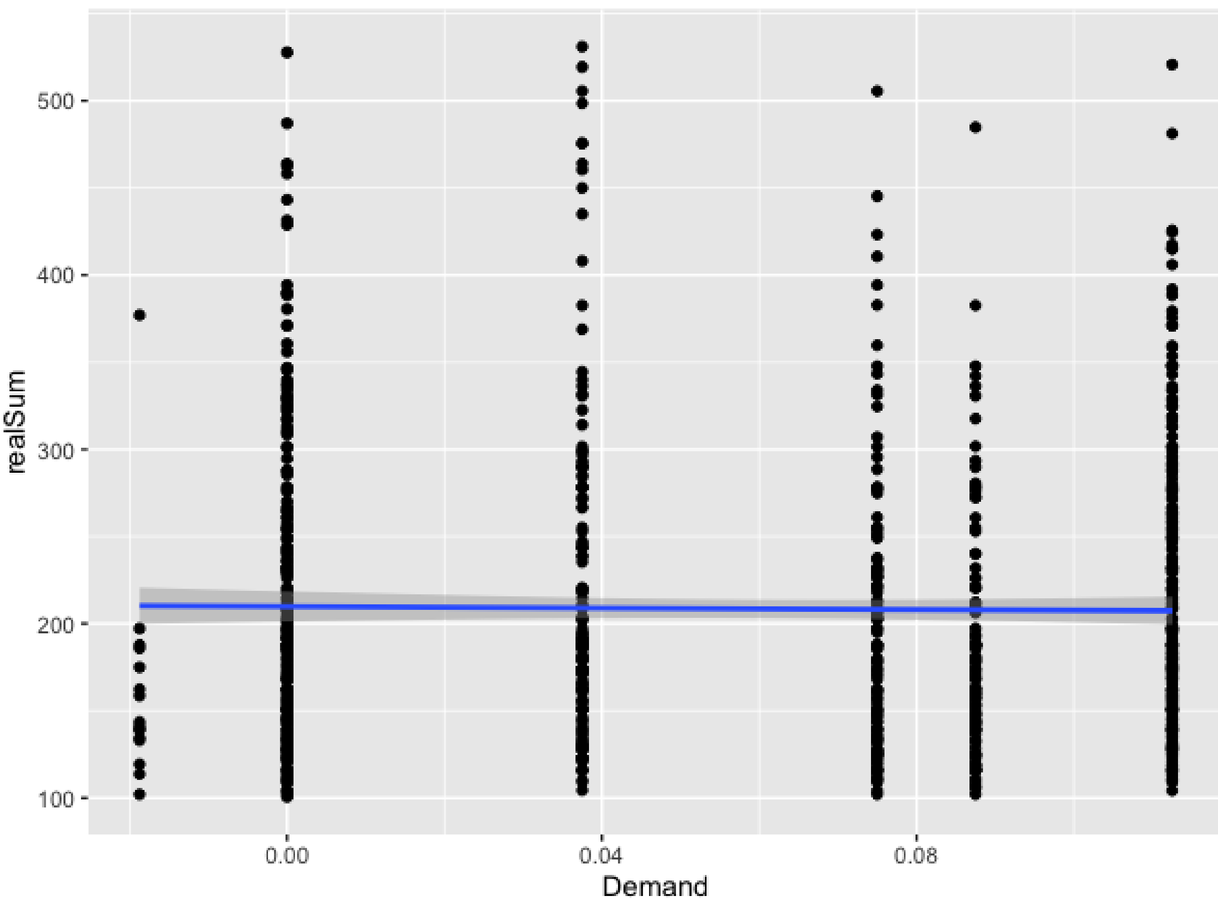
BV14



BV15



BV16



BV17

```
> cor(Berlin_c)
                              realSum Private_Room person_capacity Host_Superhost       multi
realSum                   1.000000000  -0.57639231     0.427179094   -0.004008654  0.01825272
Private_Room             -0.576392306   1.00000000    -0.399692994    0.050654608  0.05256209
person_capacity           0.427179094  -0.39969299     1.000000000   -0.039002253  0.06407967
Host_Superhost           -0.004008654   0.05065461    -0.039002253    1.000000000  0.07702071
multi                     0.018252717   0.05256209     0.064079666    0.077020709  1.00000000
biz                       0.257891635  -0.17969785     0.274849153   -0.119128976 -0.26858561
cleanliness_rating        0.063080252   0.01665548    -0.132158148    0.225376769 -0.01750314
guest_satisfaction_overall -0.313872979  0.26068366    -0.289538544    0.212734581 -0.02884115
bedrooms                  0.223294532  -0.07729069     0.353478627    0.044664515  0.05273346
dist                     -0.180808202   0.01173431    -0.011745749   -0.015243139 -0.03000032
metro_dist               -0.107753681  -0.06148747     0.045858307   -0.072945545 -0.03527729
attr_index_norm           0.290173437  -0.08442821    -0.014657229    0.031343971 -0.01241715
rest_index_norm           0.280205030  -0.08727432    -0.001091894    0.029648588  0.01000370
crime_rate                0.064148124   0.02161185    -0.030776687    0.017407843 -0.01614045
Attractions_count         0.061690418  -0.01857332    -0.040119697   -0.046799400 -0.05025109
Demand                   -0.012191246   0.05473978    -0.022546951    0.029754785  0.00457717
                                 biz cleanliness_rating guest_satisfaction_overall     bedrooms
realSum                   0.25789163        0.063080252                -0.31387298  0.223294532
Private_Room             -0.17969785        0.016655482                 0.26068366 -0.077290693
person_capacity           0.27484915       -0.132158148                -0.28953854  0.353478627
Host_Superhost           -0.11912898        0.225376769                 0.21273458  0.044664515
multi                    -0.26858561       -0.017503137                -0.02884115  0.052733462
biz                       1.00000000       -0.133139469                -0.23019308 -0.075331554
cleanliness_rating       -0.13313947        1.000000000                 0.91223782 -0.025371463
guest_satisfaction_overall -0.23019308      0.912237820                 1.00000000 -0.103460633
bedrooms                 -0.07533155       -0.025371463                -0.10346063  1.000000000
dist                     -0.08517725        0.071497401                 0.13358658  0.049778252
metro_dist               -0.11892319       -0.005762091                 0.03549805  0.071002928
attr_index_norm           0.21247639        0.035958894                -0.07114336 -0.077754442
rest_index_norm           0.21402717        0.017072908                -0.09181789 -0.081739956
crime_rate                0.04688833       -0.018545207                -0.03297036 -0.032777230
Attractions_count         0.02749896        0.038277087                 0.02575033 -0.048329302
Demand                    0.05211992       -0.017374849                -0.01401456 -0.001244438
```

CA1

	dist	metro_dist	attr_index_norm	rest_index_norm	crime_rate
realSum	-0.18080820	-0.107753681	0.29017344	0.280205030	0.06414812
Private_Room	0.01173431	-0.061487473	-0.08442821	-0.087274321	0.02161185
person_capacity	-0.01174575	0.045858307	-0.01465723	-0.001091894	-0.03077669
Host_Superhost	-0.01524314	-0.072945545	0.03134397	0.029648588	0.01740784
multi	-0.03000032	-0.035277287	-0.01241715	0.010003695	-0.01614045
biz	-0.08517725	-0.118923195	0.21247639	0.214027169	0.04688833
cleanliness_rating	0.07149740	-0.005762091	0.03595889	0.017072908	-0.01854521
guest_satisfaction_overall	0.13358658	0.035498053	-0.07114336	-0.091817887	-0.03297036
bedrooms	0.04977825	0.071002928	-0.07775444	-0.081739956	-0.03277723
dist	1.00000000	0.356174716	-0.62275187	-0.592399424	-0.27145675
metro_dist	0.35617472	1.000000000	-0.38815979	-0.433902524	-0.29988805
attr_index_norm	-0.62275187	-0.388159790	1.00000000	0.851555190	0.48602953
rest_index_norm	-0.59239942	-0.433902524	0.85155519	1.000000000	0.32286791
crime_rate	-0.27145675	-0.299888052	0.48602953	0.322867907	1.00000000
Attractions_count	-0.12169192	-0.123800647	0.41006326	0.245068842	0.50749683
Demand	0.24023851	-0.289333729	0.06105305	0.087106510	0.59042613

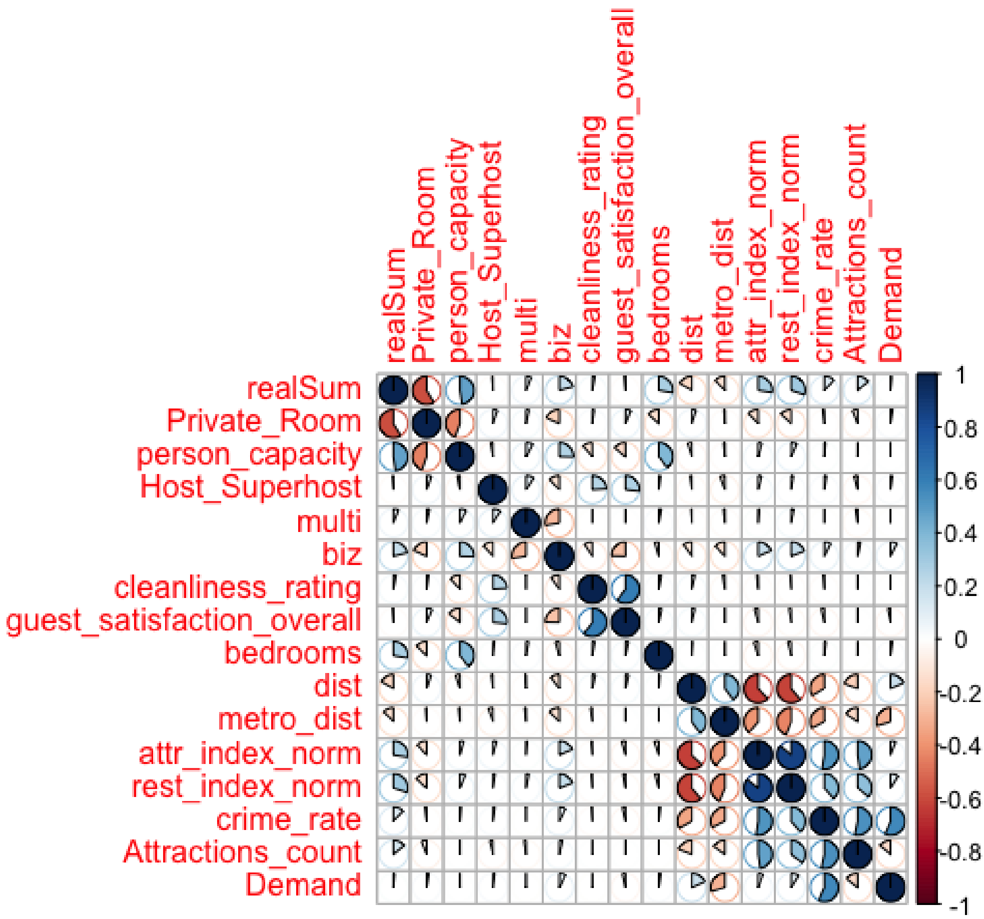
	Attractions_count	Demand
realSum	0.06169042	-0.012191246
Private_Room	-0.01857332	0.054739784
person_capacity	-0.04011970	-0.022546951
Host_Superhost	-0.04679940	0.029754785
multi	-0.05025109	0.004577170
biz	0.02749896	0.052119922
cleanliness_rating	0.03827709	-0.017374849
guest_satisfaction_overall	0.02575033	-0.014014560
bedrooms	-0.04832930	-0.001244438
dist	-0.12169192	0.240238510
metro_dist	-0.12380065	-0.289333729
attr_index_norm	0.41006326	0.061053046
rest_index_norm	0.24506884	0.087106510
crime_rate	0.50749683	0.590426132
Attractions_count	1.00000000	-0.132353016
Demand	-0.13235302	1.000000000

```
> berlin_pvalues = berlin_rcorr$P
> berlin_pvalues
                               realSum Private_Room person_capacity Host_Superhost      multi
realSum                             NA 0.000000e+00    0.0000000000   9.017445e-01 0.57398720
Private_Room              0.000000e+00           NA    0.0000000000   1.185122e-01 0.10525097
person_capacity           0.000000e+00 0.000000e+00              NA   2.295038e-01 0.04820704
Host_Superhost            9.017445e-01 1.185122e-01    0.2295037940             NA 0.01752080
multi                     5.739872e-01 1.052510e-01    0.0482070354   1.752080e-02         NA
biz                       6.661338e-16 2.408425e-08    0.0000000000   2.313922e-04 0.00000000
cleanliness_rating        5.181526e-02 6.079607e-01    0.0000434893   2.039480e-12 0.58981911
guest_satisfaction_overall 0.000000e+00 4.440892e-16   0.0000000000   3.408429e-11 0.37431122
bedrooms                  3.281819e-12 1.712818e-02    0.0000000000   1.687435e-01 0.10411958
dist                      1.967660e-08 7.177985e-01    0.7175352164   6.387256e-01 0.35540835
metro_dist                8.738918e-04 5.803040e-02    0.1576336223   2.447723e-02 0.27712582
attr_index_norm           0.000000e+00 9.191474e-03    0.6516770517   3.342642e-01 0.70213901
rest_index_norm           0.000000e+00 7.081422e-03    0.9731739089   3.610795e-01 0.75800715
crime_rate                4.796779e-02 5.056197e-01    0.3430904337   5.918461e-01 0.61910327
Attractions_count         5.720563e-02 5.672776e-01    0.2164239011   1.492726e-01 0.12147843
Demand                    7.073060e-01 9.157911e-02    0.4873792456   3.593613e-01 0.88789588
                                   biz cleanliness_rating guest_satisfaction_overall     bedrooms
realSum                   6.661338e-16       5.181526e-02               0.000000e+00 3.281819e-12
Private_Room              2.408425e-08       6.079607e-01               4.440892e-16 1.712818e-02
person_capacity           0.000000e+00       4.348930e-05               0.000000e+00 0.000000e+00
Host_Superhost            2.313922e-04       2.039480e-12               3.408429e-11 1.687435e-01
multi                     0.000000e+00       5.898191e-01               3.743112e-01 1.041196e-01
biz                                 NA       3.809455e-05               6.665779e-13 2.016075e-02
cleanliness_rating        3.809455e-05                 NA               0.000000e+00 4.345030e-01
guest_satisfaction_overall 6.665779e-13      0.000000e+00                         NA 1.398648e-03
bedrooms                  2.016075e-02       4.345030e-01               1.398648e-03           NA
dist                      8.587762e-03       2.746956e-02               3.585275e-05 1.250276e-01
metro_dist                2.372768e-04       8.591464e-01               2.741287e-01 2.856104e-02
attr_index_norm           3.603740e-11       2.679436e-01               2.824732e-02 1.647174e-02
rest_index_norm           2.575939e-11       5.989958e-01               4.600239e-03 1.168139e-02
crime_rate                1.485003e-01       5.678645e-01               3.097779e-01 3.126220e-01
Attractions_count         3.969586e-01       2.382869e-01               4.276711e-01 1.364076e-01
Demand                    1.082152e-01       5.925487e-01               6.660047e-01 9.694278e-01
```

P1

```
                                    dist   metro_dist attr_index_norm rest_index_norm crime_rate
realSum                     1.967660e-08 0.0008738918    0.000000e+00    0.000000e+00 0.04796779
Private_Room                7.177985e-01 0.0580303970    9.191474e-03    7.081422e-03 0.50561973
person_capacity             7.175352e-01 0.1576336223    6.516771e-01    9.731739e-01 0.34309043
Host_Superhost              6.387256e-01 0.0244772348    3.342642e-01    3.610795e-01 0.59184614
multi                       3.554084e-01 0.2771258193    7.021390e-01    7.580072e-01 0.61910327
biz                         8.587762e-03 0.0002372768    3.603740e-11    2.575939e-11 0.14850034
cleanliness_rating          2.746956e-02 0.8591464481    2.679436e-01    5.989958e-01 0.56786449
guest_satisfaction_overall  3.585275e-05 0.2741286540    2.824732e-02    4.600239e-03 0.30977789
bedrooms                    1.250276e-01 0.0285610416    1.647174e-02    1.168139e-02 0.31262200
dist                                  NA 0.0000000000    0.000000e+00    0.000000e+00 0.00000000
metro_dist                  0.000000e+00           NA    0.000000e+00    0.000000e+00 0.00000000
attr_index_norm             0.000000e+00 0.0000000000              NA    0.000000e+00 0.00000000
rest_index_norm             0.000000e+00 0.0000000000    0.000000e+00              NA 0.00000000
crime_rate                  0.000000e+00 0.0000000000    0.000000e+00    0.000000e+00         NA
Attractions_count           1.686789e-04 0.0001294451    0.000000e+00    1.776357e-14 0.00000000
Demand                      5.950795e-14 0.0000000000    5.982908e-02    7.192599e-03 0.00000000
                            Attractions_count       Demand
realSum                          5.720563e-02 7.073060e-01
Private_Room                     5.672776e-01 9.157911e-02
person_capacity                  2.164239e-01 4.873792e-01
Host_Superhost                   1.492726e-01 3.593613e-01
multi                            1.214784e-01 8.878959e-01
biz                              3.969586e-01 1.082152e-01
cleanliness_rating               2.382869e-01 5.925487e-01
guest_satisfaction_overall       4.276711e-01 6.660047e-01
bedrooms                         1.364076e-01 9.694278e-01
dist                             1.686789e-04 5.950795e-14
metro_dist                       1.294451e-04 0.000000e+00
attr_index_norm                  0.000000e+00 5.982908e-02
rest_index_norm                  1.776357e-14 7.192599e-03
crime_rate                       0.000000e+00 0.000000e+00
Attractions_count                          NA 4.236353e-05
Demand                           4.236353e-05           NA
```

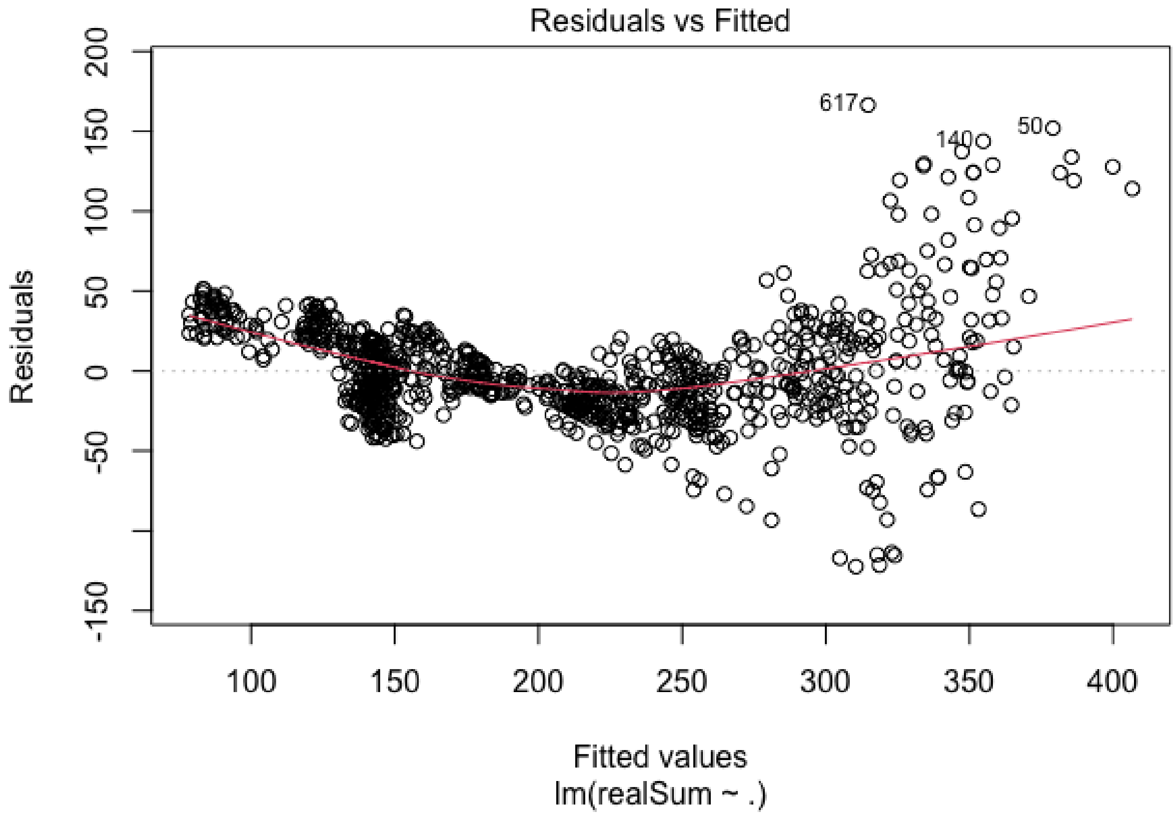
P2


realSum
Private_Room
person_capacity
Host_Superhost
multi
biz
cleanliness_rating
guest_satisfaction_overall
bedrooms
dist
metro_dist
attr_index_norm
rest_index_norm
crime_rate
Attractions_count
Demand
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

CM2

```
Call:
lm(formula = Berlin_c$realSum ~ ., data = Berlin_c)

Residuals:
     Min       1Q   Median       3Q      Max
-122.362  -19.424   -3.943   18.076  166.491

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 950.8036    29.6720  32.044  < 2e-16 ***
Private_Room                 -8.5616     3.0178  -2.837 0.004651 **
person_capacity               4.7518     1.2832   3.703 0.000225 ***
Host_Superhost               -0.5980     2.5409  -0.235 0.813992
multi                        -1.0838     2.6338  -0.411 0.680813
biz                           2.1405     3.4250   0.625 0.532140
cleanliness_rating          264.1709     5.8335  45.285  < 2e-16 ***
guest_satisfaction_overall -34.6605     0.7671 -45.183  < 2e-16 ***
bedrooms                      9.4062     2.9764   3.160 0.001627 **
dist                         -0.2024     0.9012  -0.225 0.822363
metro_dist                   -1.0115     2.1784  -0.464 0.642519
attr_index_norm               1.0959     0.3955   2.771 0.005701 **
rest_index_norm              -0.1449     0.2175  -0.666 0.505494
crime_rate                   -0.2439     0.7556  -0.323 0.746929
Attractions_count             0.1472     0.2168   0.679 0.497410
Demand                        3.3702    64.9426   0.052 0.958624
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.52 on 935 degrees of freedom
Multiple R-squared:  0.8387,    Adjusted R-squared:  0.8361
F-statistic: 324.1 on 15 and 935 DF,  p-value: < 2.2e-16
```

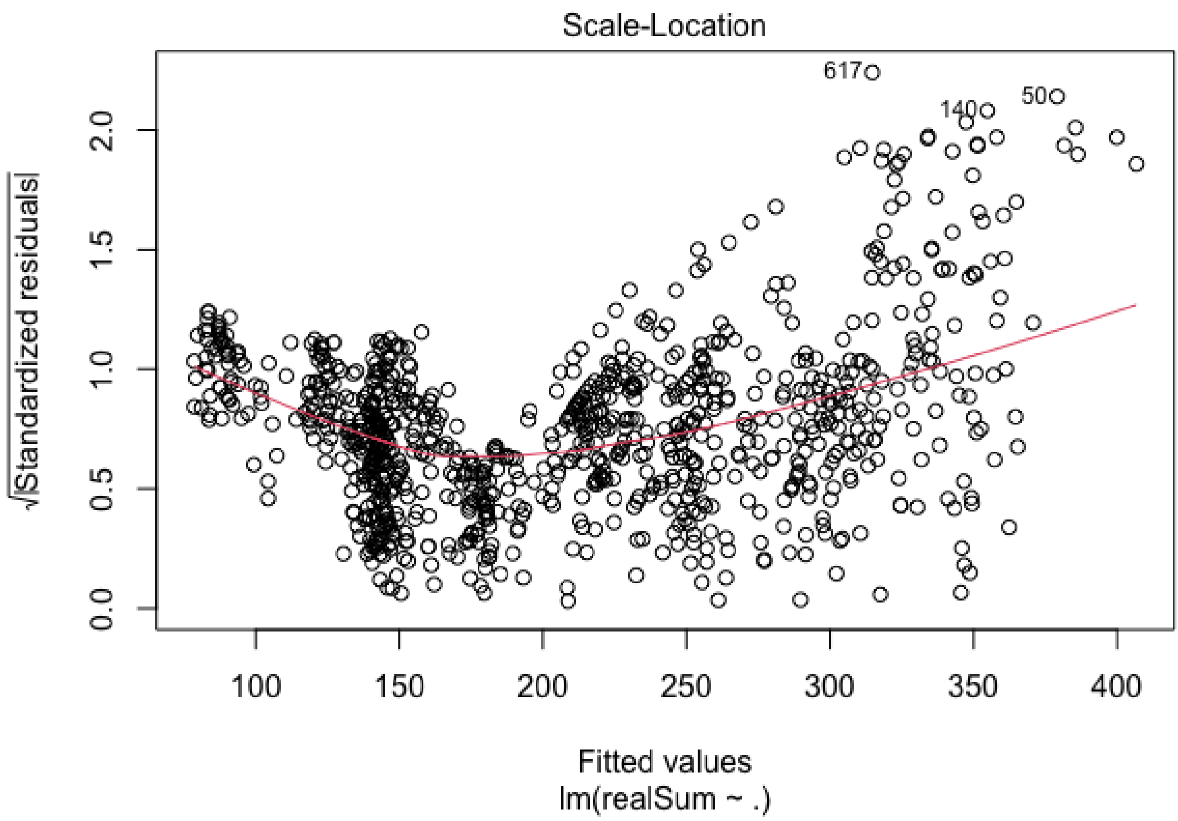
MS 1



MS 2

```
> durbinWatsonTest(LM)
 lag Autocorrelation D-W Statistic p-value
   1     -0.01545507      2.023868    0.77
 Alternative hypothesis: rho != 0
```

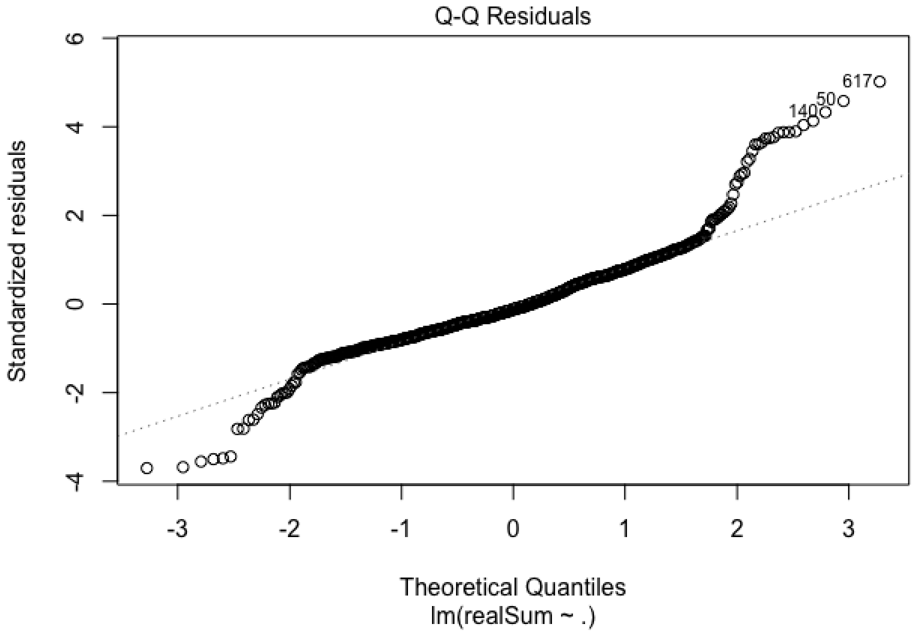
MS 3



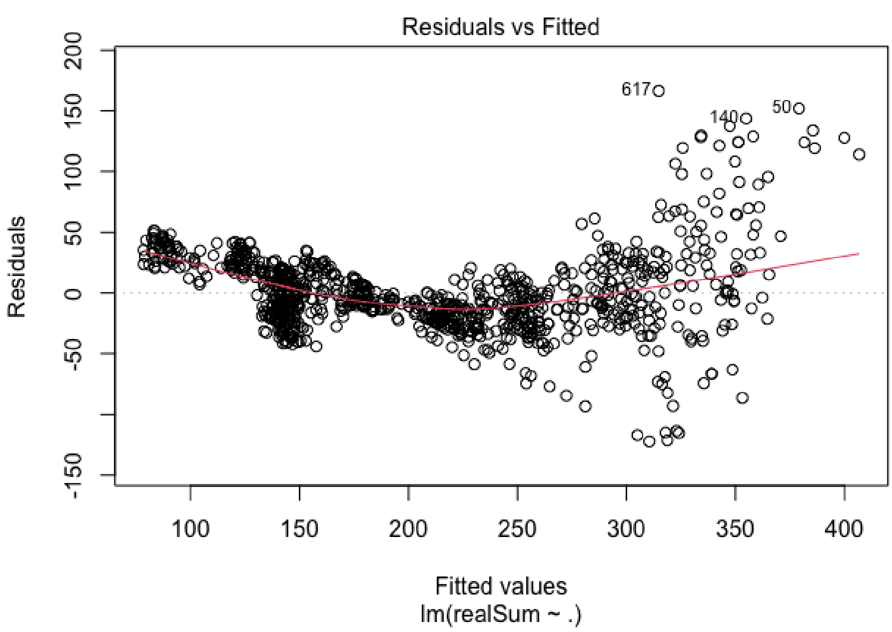
MS 4

```
> ncvTest(LM)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 398.6441, Df = 1, p = < 2.22e-16
```

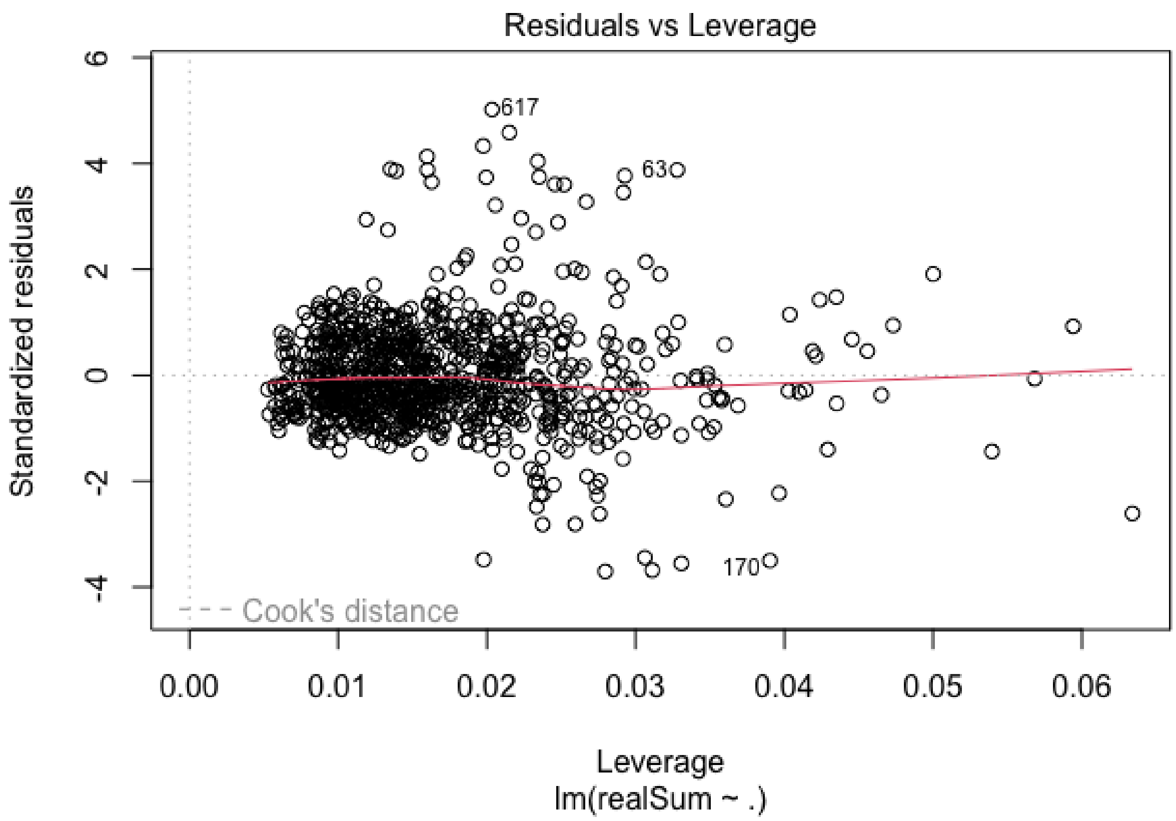
MS 5


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(realSum ~ .)
617
50
140
6
4
2
0
-2
-4
-3
-2
-1
0
1
2
3

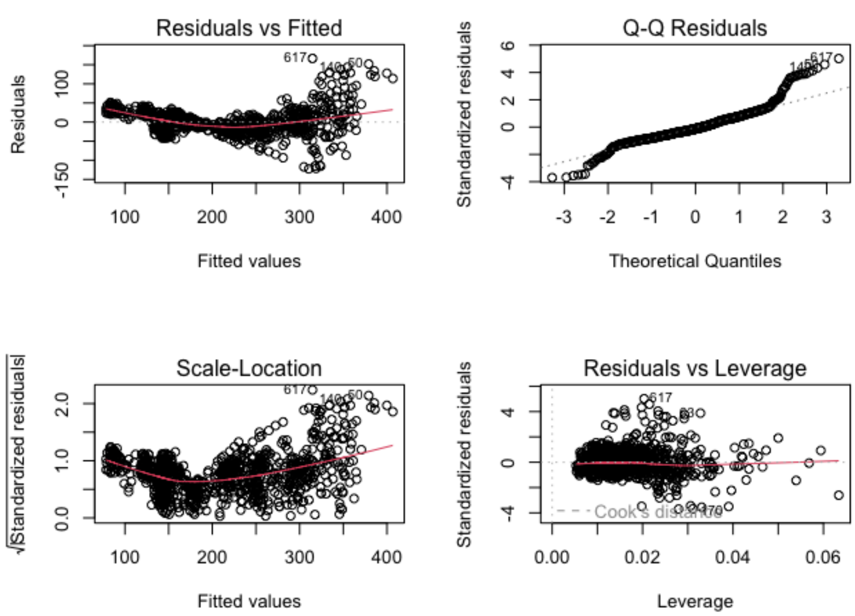
MS 6


Residuals vs Fitted
Residuals
Fitted values
lm(realSum ~ .)
617
140
50
200
150
100
50
0
-50
-150
100
150
200
250
300
350
400

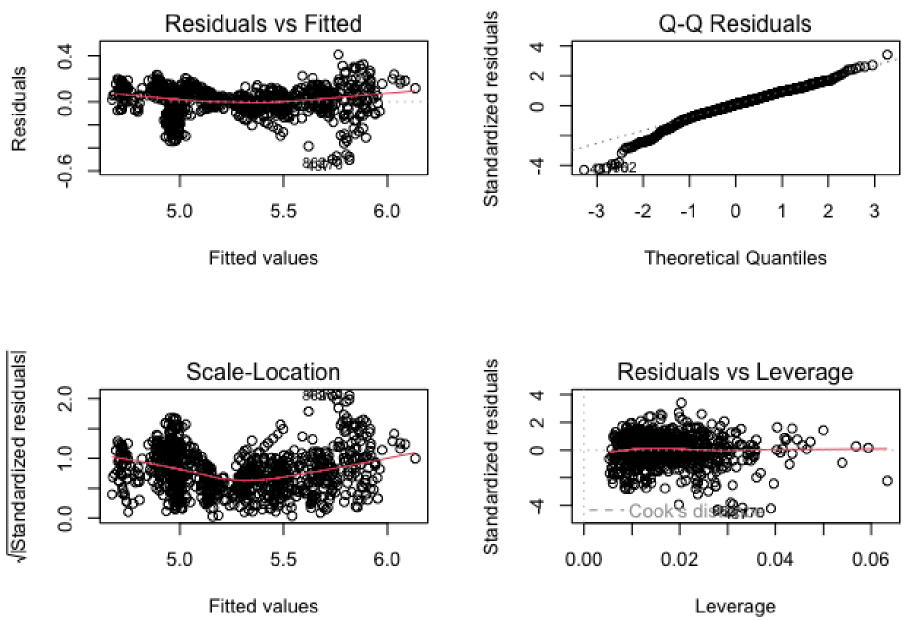
MS 7


Residuals vs Leverage
Standardized residuals
Leverage
lm(realSum ~ .)
Cook's distance
617
63
170

MS 8


Residuals vs Fitted
Residuals
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

DT 1

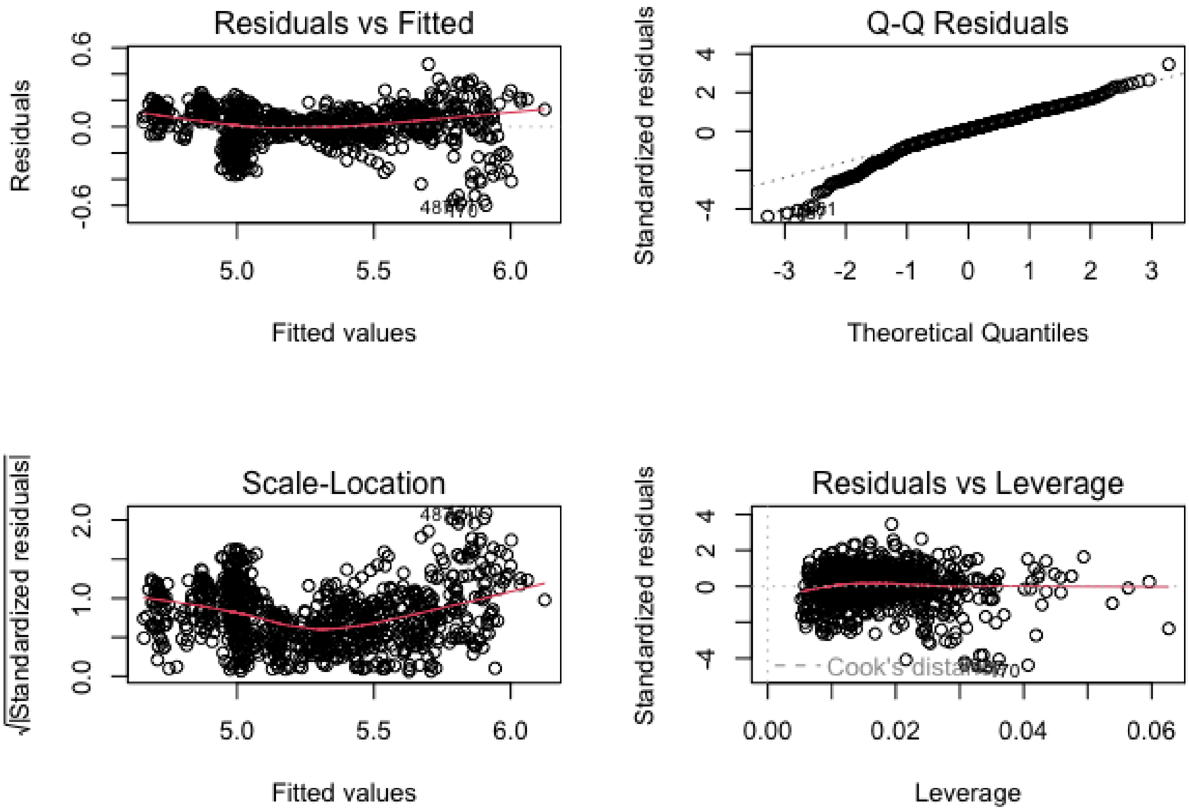


Residuals vs Fitted
Residuals
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

DT 2

```
> summary(lm1)

Call:
lm(formula = realSum ~ ., data = Berlin_c)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.51739 -0.06153  0.01083  0.07991  0.41216 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                8.795e+00  1.080e-01  81.441  < 2e-16 ***
Private_Room              -4.508e-02  1.098e-02  -4.105  4.4e-05 ***
person_capacity            1.499e-02  4.670e-03   3.210  0.00137 ** 
Host_Superhost             8.901e-04  9.247e-03   0.096  0.92334    
multi                      1.142e-02  9.585e-03   1.191  0.23376    
biz                        1.331e-02  1.246e-02   1.068  0.28582    
cleanliness_rating         1.200e+00  2.123e-02  56.523  < 2e-16 ***
guest_satisfaction_overall -1.584e-01  2.792e-03 -56.755  < 2e-16 ***
bedrooms                   1.569e-02  1.083e-02   1.448  0.14784    
dist                      -3.843e-03  3.280e-03  -1.172  0.24157    
metro_dist                 3.302e-03  7.928e-03   0.417  0.67711    
attr_index_norm            2.790e-03  1.439e-03   1.938  0.05288 .  
rest_index_norm            1.416e-05  7.917e-04   0.018  0.98573    
crime_rate                -1.091e-03  2.750e-03  -0.397  0.69159    
Attractions_count          7.867e-04  7.891e-04   0.997  0.31904    
Demand                     1.862e-01  2.363e-01   0.788  0.43099    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.122 on 935 degrees of freedom
Multiple R-squared:  0.8885,	Adjusted R-squared:  0.8867 
F-statistic: 496.7 on 15 and 935 DF,  p-value: < 2.2e-16
```

DT 3

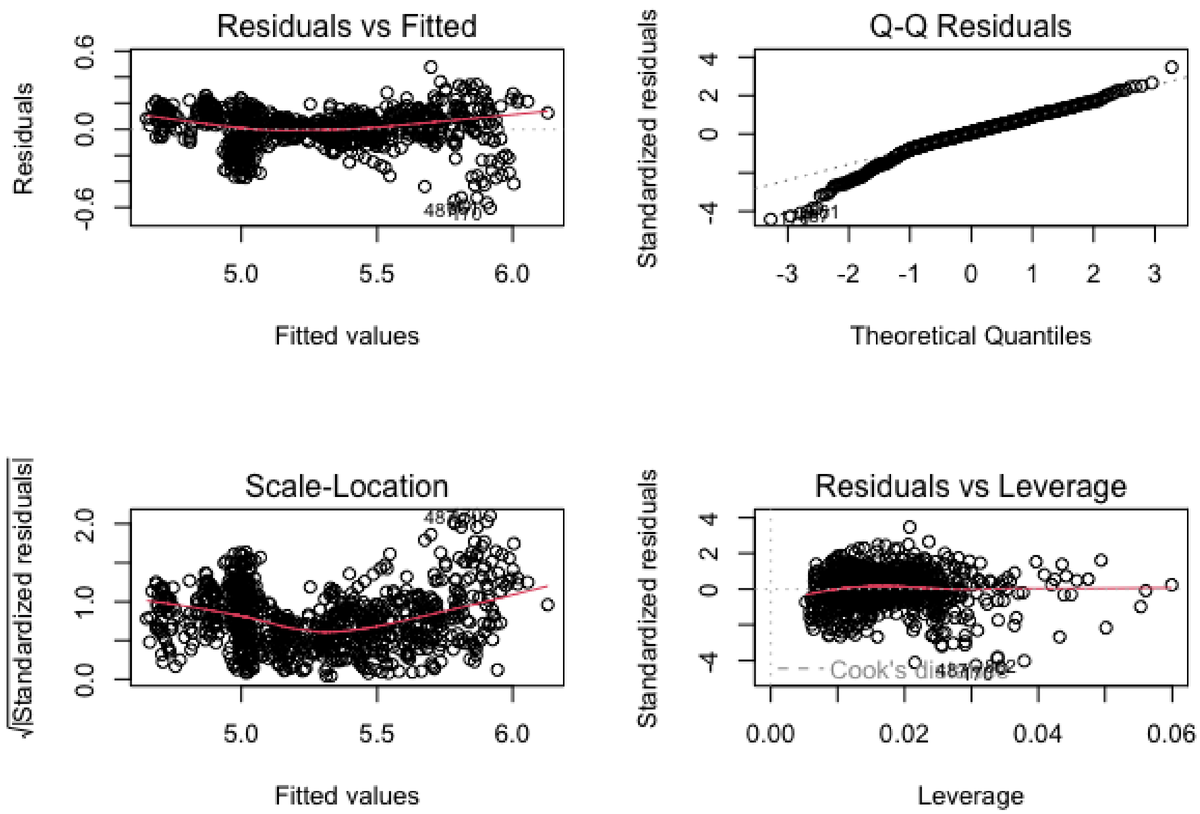


Residuals vs Fitted
Residuals
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

DT 4

```
Call:
lm(formula = realSum ~ ., data = Berlin_c)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.60467 -0.06656  0.01213  0.08585  0.47753 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 5.683e+01  1.111e+00  51.162  < 2e-16 ***
Private_Room               -7.009e-02  1.241e-02  -5.648 2.15e-08 ***
person_capacity             1.828e-02  5.331e-03   3.429 0.000632 ***
Host_Superhost             -9.463e-04  1.055e-02  -0.090 0.928577    
multi                       1.873e-02  1.092e-02   1.715 0.086630 .  
biz                         2.288e-02  1.417e-02   1.615 0.106726    
cleanliness_rating          1.138e+00  2.386e-02  47.696  < 2e-16 ***
guest_satisfaction_overall -1.376e+01  2.884e-01 -47.729  < 2e-16 ***
bedrooms                    1.917e-02  1.237e-02   1.550 0.121418    
dist                       -6.001e-04  4.375e-03  -0.137 0.890930    
metro_dist                  4.455e-03  9.073e-03   0.491 0.623551    
attr_index_norm             5.892e-02  3.678e-02   1.602 0.109560    
rest_index_norm             1.030e-02  3.326e-02   0.310 0.756871    
crime_rate                 -8.011e-05  3.229e-03  -0.025 0.980215    
Attractions_count           4.686e-04  9.369e-04   0.500 0.617127    
Demand                      8.546e-02  2.851e-01   0.300 0.764383    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1391 on 935 degrees of freedom
Multiple R-squared:  0.855,     Adjusted R-squared:  0.8526 
F-statistic: 367.5 on 15 and 935 DF,  p-value: < 2.2e-16
```

DT 5


Residuals vs Fitted
Residuals
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

DT 6

```
Call:
lm(formula = train_set$realSum ~ ., data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.45418 -0.06074  0.00279  0.07373  0.32610 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 8.6071393  0.1328428  64.792  < 2e-16 ***
Private_Room               -0.0436155  0.0137898  -3.163  0.00164 ** 
person_capacity             0.0199274  0.0057437   3.469  0.00056 ***
Host_Superhost             -0.0030383  0.0112089  -0.271  0.78644    
multi                       0.0128133  0.0116780   1.097  0.27300    
biz                         0.0061073  0.0150985   0.404  0.68599    
cleanliness_rating          1.2040980  0.0264147  45.584  < 2e-16 ***
guest_satisfaction_overall -0.1562000  0.0034765 -44.931  < 2e-16 ***
bedrooms                    0.0119696  0.0134049   0.893  0.37226    
dist                       -0.0106920  0.0060258  -1.774  0.07652 .  
metro_dist                 -0.0437381  0.0175262  -2.496  0.01285 *  
attr_index_norm             0.0029106  0.0015134   1.923  0.05494 .  
rest_index_norm            -0.0004867  0.0009063  -0.537  0.59151    
crime_rate                 -0.0010825  0.0040031  -0.270  0.78694    
Attractions_count           0.0003277  0.0011288   0.290  0.77167    
Demand                      0.0315754  0.3815029   0.083  0.93407    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.118 on 584 degrees of freedom
Multiple R-squared:  0.8964,    Adjusted R-squared:  0.8938 
F-statistic:   337 on 15 and 584 DF,  p-value: < 2.2e-16
```

MV 1

```
Call:
lm(formula = train_set$realSum ~ Private_Room + person_capacity +
    cleanliness_rating + guest_satisfaction_overall + dist +
    metro_dist + attr_index_norm, data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46360 -0.06149  0.00397  0.07613  0.33740

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.6251895  0.1168732  73.800  < 2e-16 ***
Private_Room              -0.0402485  0.0133654  -3.011  0.00271 **
person_capacity            0.0216753  0.0052643   4.117 4.38e-05 ***
cleanliness_rating         1.2095059  0.0252158  47.966  < 2e-16 ***
guest_satisfaction_overall -0.1570657  0.0032954 -47.662  < 2e-16 ***
dist                      -0.0105269  0.0033724  -3.121  0.00189 **
metro_dist                -0.0403450  0.0168101  -2.400  0.01670 *
attr_index_norm            0.0024282  0.0009423   2.577  0.01021 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1174 on 592 degrees of freedom
Multiple R-squared:  0.896,     Adjusted R-squared:  0.8948
F-statistic: 728.7 on 7 and 592 DF,  p-value: < 2.2e-16
```

MV 2

	OLS MODEL	
	Overall Data	
Mape	Mdape	MSE
0.01876289	0.0143836	0.01626154
1.88%	1.44%	1.63%
	TEST SET	
Mape	Mdape	MSE
0.02237697	0.01882502	0.02089353
2.24%	1.88%	2.09%

OLS 1

	STEPWISE	
	Overall Data	
Mape	Mdape	MSE
0.09822696	0.07707476	0.01633755
9.82%	7.71%	1.63%
	TEST SET	
Mape	Mdape	MSE
0.1157554	0.1008786	0.02100255
11.60%	10.10%	2.10%

ST 1