describe(cities)

vars n mean sd median

CityName\* 1 13232 NaN NA NA

Population 2 13232 4416836.87 4258386.00 3046163.0

CityRank 3 13232 14.83 13.51 9.0

IsMetroCity 4 13232 0.28 0.45 0.0

IsTouristDestination 5 13232 0.70 0.46 1.0

IsWeekend 6 13232 0.62 0.48 1.0

IsNewYearEve 7 13232 0.12 0.33 0.0

Date\* 8 13232 NaN NA NA

HotelName\* 9 13232 NaN NA NA

RoomRent 10 13232 5473.99 7333.12 4000.0

StarRating 11 13232 3.46 0.76 3.0

Airport 12 13232 21.16 22.76 15.0

HotelAddress\* 13 13232 NaN NA NA

HotelPincode 14 13232 397430.26 259837.50 395003.0

HotelDescription\* 15 13224 3.46 0.48 3.5

FreeWifi 16 13232 0.93 0.26 1.0

FreeBreakfast 17 13232 0.65 0.48 1.0

HotelCapacity 18 13232 62.51 76.66 34.0

HasSwimmingPool 19 13232 0.36 0.48 0.0

trimmed mad min max range

CityName\* NaN NA Inf -Inf -Inf

Population 4040816.22 3846498.95 8096.0 12442373 12434277.0

CityRank 13.30 11.86 0.0 44 44.0

IsMetroCity 0.23 0.00 0.0 1 1.0

IsTouristDestination 0.75 0.00 0.0 1 1.0

IsWeekend 0.65 0.00 0.0 1 1.0

IsNewYearEve 0.03 0.00 0.0 1 1.0

Date\* NaN NA Inf -Inf -Inf

HotelName\* NaN NA Inf -Inf -Inf

RoomRent 4383.33 2653.85 299.0 322500 322201.0

StarRating 3.40 0.74 0.0 5 5.0

Airport 16.39 11.12 0.2 124 123.8

HotelAddress\* NaN NA Inf -Inf -Inf

HotelPincode 388540.47 257975.37 100025.0 7000157 6900132.0

HotelDescription\* 3.41 0.74 3.0 5 2.0

FreeWifi 1.00 0.00 0.0 1 1.0

FreeBreakfast 0.69 0.00 0.0 1 1.0

HotelCapacity 46.03 28.17 0.0 600 600.0

HasSwimmingPool 0.32 0.00 0.0 1 1.0

skew kurtosis se

CityName\* NA NA NA

Population 0.68 -1.08 37019.65

CityRank 0.69 -0.76 0.12

IsMetroCity 0.96 -1.08 0.00

IsTouristDestination -0.86 -1.26 0.00

IsWeekend -0.51 -1.74 0.00

IsNewYearEve 2.28 3.18 0.00

Date\* NA NA NA

HotelName\* NA NA NA

RoomRent 16.75 582.06 63.75

StarRating 0.48 0.25 0.01

Airport 2.73 7.89 0.20

HotelAddress\* NA NA NA

HotelPincode 9.99 249.76 2258.86

HotelDescription\* 0.93 0.78 0.00

FreeWifi -3.25 8.57 0.00

FreeBreakfast -0.62 -1.61 0.00

HotelCapacity 2.95 11.39 0.67

HasSwimmingPool 0.60 -1.64 0.00

Our dependent variable is the Room rent.

We have a combination of independent variables(categorical and continuous) for the above dependent variable.

**Independent categorical variables.**

1. IsMetroCity

First we use the table command to find the number of cities who are metro cities and the number of cities who are not metro cities.

table(cities$IsMetroCity)

0 1

9472 3760

We use the aggregate command to find the mean Room Rent of the metro cities and the non-metro cities. There is a significant difference who gives us an idea that IsMetroCity is playing a part in determining the Room rent for a particular city.

aggregate(cities$RoomRent, by=list(cities$IsMetroCity), mean)

Group.1 x

0 5782.794

1 4696.073

We use the t test to find out if the Room rent for the metro cities and non-metro cities is the same.

H0: The room rent for metro cities and non-metro cities is same

H1:The room rent for metro cities and non metro cities is NOT same.

A very low p value helps us to reject the NULL hypothesis and conclude that the room rent for metro cities and non metro cities is not the same. Clearly the variable IsMetroCity is playing a role in determining room rent.

t.test(RoomRent~IsMetroCity, data=cities)

Welch Two Sample t-test

data: RoomRent by IsMetroCity

t = 10.721, df = 13224, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

888.0308 1285.4102

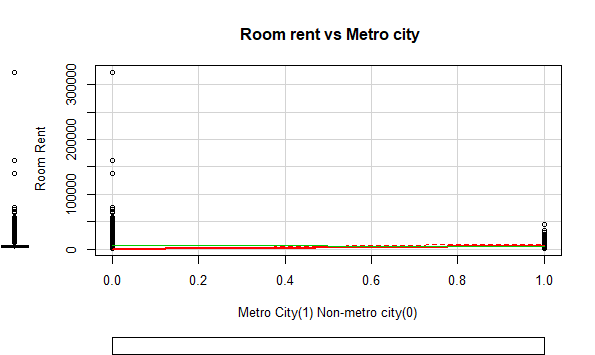
sample estimates:

mean in group 0 mean in group 1

5782.794 4696.073

We now draw a scatterplot to visually see the relationship pair-wise

scatterplot(cities$IsMetroCity, cities$RoomRent, xlab= "Metro City(1) Non-metro city(0)", ylab=" Room Rent", main=" Room rent vs Metro city")



1. IsTouristDestination

To find the number of cities who are tourist destinations and the number of cities who are not tourist destinations

table(cities$IsTouristDestination)

0 1

4007 9225

We use the aggregate command to find the mean Room Rent of the tourist destination cities and the non-tourist destination cities. There is a significant difference who gives us an idea that IsTouristDestination is playing a part in determining the Room rent for a particular city.

aggregate(cities$RoomRent, by=list(cities$IsTouristDestination), mean)

Group.1 x

1 0 4111.003

2 1 6066.024

We use the t test to find out if the Room rent for the tourist cities and non-tourist cities is the same.

H0: The room rent for tourist cities and non-tourist cities is same

H1:The room rent for tourist cities and non tourist cities is NOT same.

A very low p value helps us to reject the NULL hypothesis.

Clearly the variable IsTouristDestination is playing a role in determining room rent.

t.test(RoomRent~IsTouristDestination, data=cities)

Welch Two Sample t-test

data: RoomRent by IsTouristDestination

t = -19.449, df = 12888, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2152.059 -1757.983

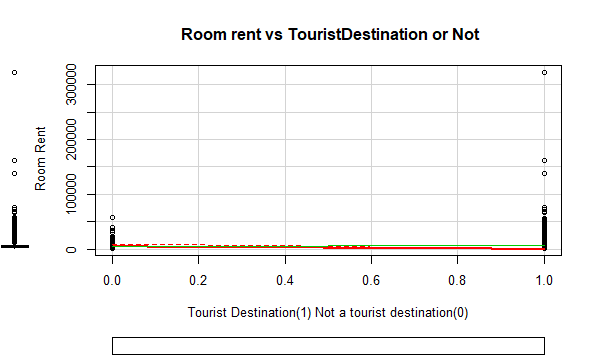
sample estimates:

mean in group 0 mean in group 1

4111.003 6066.024

A scatterplot is drawn to visually see the relationship pair-wise

scatterplot(cities$IsTouristDestination, cities$RoomRent, xlab= "Tourist Destination(1) Not a tourist destination(0)", ylab=" Room Rent", main=" Room rent vs TouristDestination or Not")



1. HasSwimmingPool

To see how many hotels have swimming pools and how many hotels do not have one.

table(cities$HasSwimmingPool)

0 1

8524 4708

We use the aggregate command to find the mean Room Rent of the hotels having swimming pools and hotels that do not have swimming pools. There is a significant difference who gives us an idea that HasSwimmingPool is playing a part in determining the Room rent for a hotel

aggregate(cities$RoomRent, by=list(cities$HasSwimmingPool), mean)

Group.1 x

1 0 3775.566

2 1 8549.052

We use the t test to find out if the Room rent for the hotels having swimming pools and the hotels that do not have swimming pools is the same.

H0: The room rent for hotels with or without swimming pool is same

H1:The room rent for hotels with and without swimming pools is NOT same.

A very low p value helps us to reject the NULL hypothesis.

Clearly the variable HasSwimmingPool is playing a role in determining room rent.

t.test(RoomRent~HasSwimmingPool, data=cities)

Welch Two Sample t-test

data: RoomRent by HasSwimmingPool

t = -29.013, df = 5011.3, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-5096.030 -4450.942

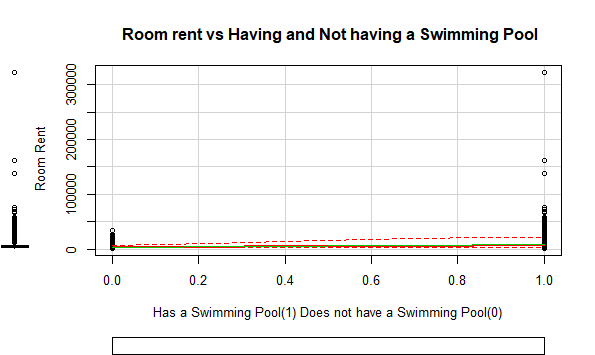
sample estimates:

mean in group 0 mean in group 1

3775.566 8549.052

A scatter plot is drawn to visually see the relationship pair-wise

scatterplot(cities$HasSwimmingPool, cities$RoomRent, xlab= "Has a Swimming Pool(1) Does not have a Swimming Pool(0)", ylab=" Room Rent", main=" Room rent vs Having and Not having a Swimming Pool")



To see the number of days which were New Year Eve

table(cities$IsNewYearEve)

0 1

11586 1646

Using the aggregate function to see the difference in the mean of the Hotel Rent on the days which were New year eve and the days which were not new year eve. There is a significant difference which gives us an idea that IsNewYearEve is playing a role in determining the Room rent

aggregate(cities$RoomRent, by=list(cities$IsNewYearEve), mean)

Group.1 x

1 0 5367.606

2 1 6222.826

We use the t test to find out if the Room rent for the hotels on New Year eve and on other days is the same.

H0: The room rent for hotels on all the days is same

H1:The room rent for hotels on New year Eve and other days is NOT same.

A very low p value helps us to reject the NULL hypothesis.

Clearly the variable IsNewYearEve is playing a role in determining room rent.

t.test(RoomRent~IsNewYearEve, data=cities)

Welch Two Sample t-test

data: RoomRent by IsNewYearEve

t = -4.1793, df = 2065, p-value = 3.046e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1256.5297 -453.9099

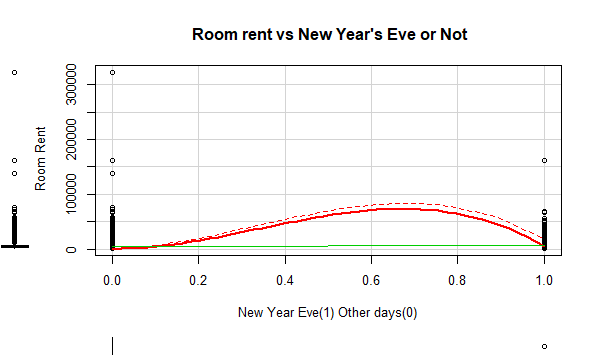
sample estimates:

mean in group 0 mean in group 1

5367.606 6222.826

A scatter plot is used to visually see the relationship pair wise.

scatterplot(cities$IsNewYearEve, cities$RoomRent, xlab= "New Year Eve(1) Other days(0)", ylab=" Room Rent", main=" Room rent vs New Year's Eve or Not")



**Independent continuous variables**

1. Star Rating

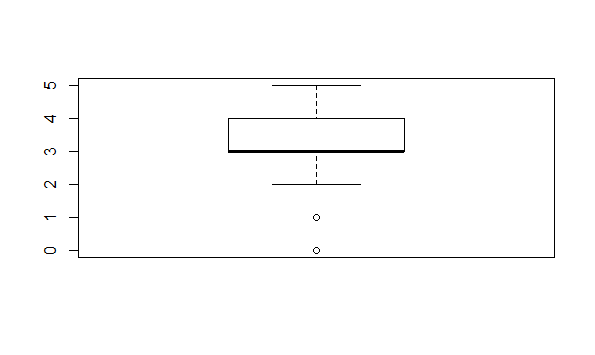
We use the cor function to see the correlation between the room rent and the star rating

cor(cities$RoomRent, cities$StarRating)

[1] 0.3693734

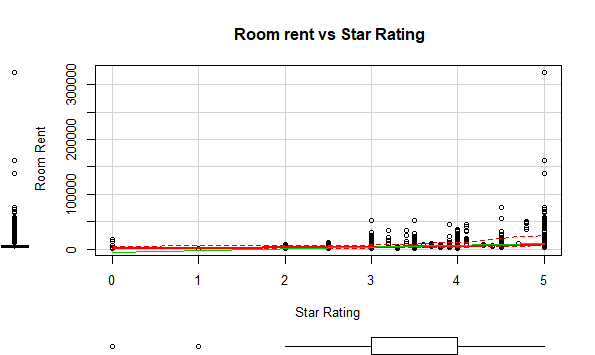
To examine the variable Star rating in detail, we use the box plot. We can see the median and the ouliers.

boxplot(cities$StarRating)



We use a scatter plot to visually see the relationship pair-wise

scatterplot(cities$StarRating, cities$RoomRent, xlab= "Star Rating", ylab=" Room Rent", main=" Room rent vs Star Rating")



1. Hotel capacity

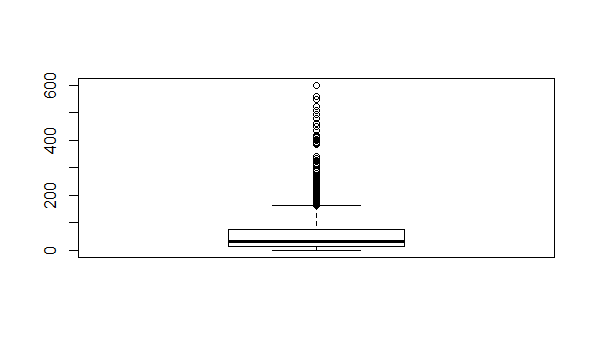
We measure the correlation using the cor command.

cor(cities$RoomRent, cities$HotelCapacity)

[1] 0.1578733

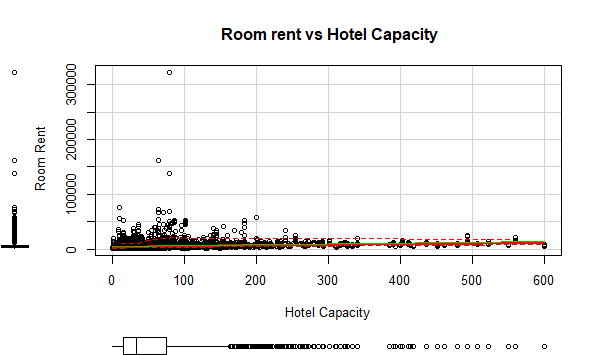
To examine the variable HotelCapacity in detail, we use boxplots.

boxplot(cities$HotelCapacity)



We use the scatterplot commands to visually see the relationship pair-wise.

scatterplot(cities$HotelCapacity, cities$RoomRent, xlab= "Hotel Capacity", ylab=" Room Rent", main=" Room rent vs Hotel Capacity")



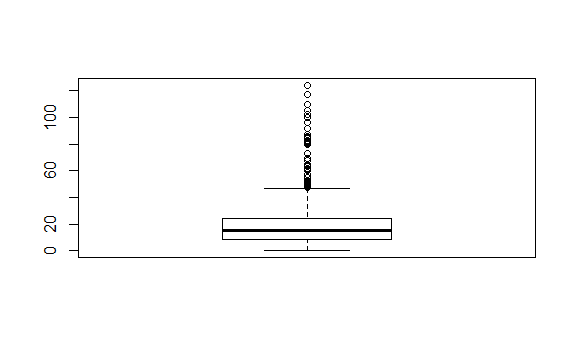
1. Airport

cor(cities$RoomRent, cities$Airport)

[1] 0.04965324

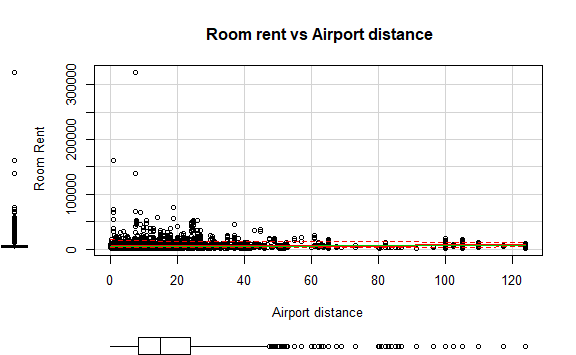
To examine the variable Airport in detail, we use boxplots.

boxplot(cities$HotelCapacity)



We use the scatterplot commands to visually see the relationship pair-wise.

scatterplot(cities$Airport, cities$RoomRent, xlab= "Airport distance", ylab=" Room Rent", main=" Room rent vs Airport distance")



Till this point we have seen 4 categorical variables and 3 continuous variables which influences Room rent(our dependent variable). The other three variables have relatively high p values(more than 0.05) and a lesser correlation.

***We select HotelCapacity, HasSwimmingPool and StarRating as our three most important independent variables, after seeing their p values and the correlation.***

**CORRGRAM**

We construct a corrgram for the dependent variable along with the three independent variables.

library(Hmisc)

library(car)

library(corrgram)

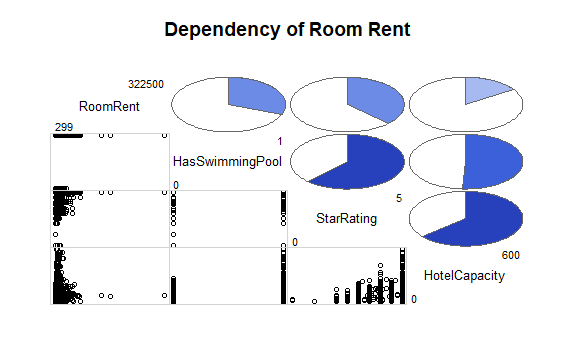
colcities <- c("RoomRent","StarRating","HotelCapacity","HasSwimmingPool")

corrgram(cities[,colcities], order=TRUE,

+ main="Dependency of RoomRent",

+ lower.panel=panel.pts, upper.panel=panel.pie,

+ diag.panel=panel.minmax, text.panel=panel.txt)



**VARIANCE-COVARIANCE MATRIX**

vcov(lm(cities$RoomRent~cities$StarRating+cities$HotelCapacity + cities$HasSwimmingPool))

(Intercept) cities$StarRating cities$HotelCapacity cities$HasSwimmingPool

(Intercept) 115973.91 -37005.33340 132.820048 20026.05852

cities$StarRating -37005.33 12470.18238 -53.584269 -7809.21320

cities$HotelCapacity 132.82 -53.58427 1.011859 -30.15313

cities$HasSwimmingPool 20026.06 -7809.21320 -30.153132 24930.59316

**CORELATION MATRIX**

We construct a correlation matrix for the dependent variable along with the three independent variables.

|  |
| --- |
| library(Hmisc)  Loading required package: lattice  Loading required package: survival  Loading required package: Formula  Loading required package: ggplot2  Attaching package: ‘ggplot2’  The following objects are masked from ‘package:psych’:  %+%, alpha  Attaching package: ‘Hmisc’  The following object is masked from ‘package:psych’:  describe  The following objects are masked from ‘package:base’:  format.pval, round.POSIXt, trunc.POSIXt, units  > colcities <- c("RoomRent","StarRating","HotelCapacity","HasSwimmingPool")  > corMatrix <- rcorr(as.matrix(cities[,colcities]))  > corMatrix  RoomRent StarRating HotelCapacity HasSwimmingPool  RoomRent 1.00 0.37 0.16 0.31  StarRating 0.37 1.00 0.64 0.62  HotelCapacity 0.16 0.64 1.00 0.51  HasSwimmingPool 0.31 0.62 0.51 1.00  n= 13232  P  RoomRent StarRating HotelCapacity HasSwimmingPool  RoomRent 0 0 0  StarRating 0 0 0  HotelCapacity 0 0 0  HasSwimmingPool 0 0 0 |
|  |
| |  | | --- | |  | |

**T tests**

We run t tests to see dependency of Roomrent on all the categorical variables.

1)

H0: The room rent of hotels on weekends and weekdays are the same

H1: The room rent of hotels on weekends are more than on weekdays.

aggregate(cities$RoomRent, by=list(cities$IsWeekend), mean)

Group.1 x

1 0 5430.835

2 1 5500.129

t.test(RoomRent~IsWeekend, data=cities)

Welch Two Sample t-test

data: RoomRent by IsWeekend

t = -0.51853, df = 9999.4, p-value = 0.6041

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-331.2427 192.6559

sample estimates:

mean in group 0 mean in group 1

5430.835 5500.129

As we can see the difference between the means of prices on weekends and weekdays are almost similar.

P value is greater than 0.05 so we fail to reject the null hypothesis.

2)

H0: The room rent for hotels on New year eve and on other days is the same

H1: The room rent for hotels on New year eve is more than the room rent on the other days.

aggregate(cities$RoomRent, by=list(cities$IsNewYearEve), mean)

Group.1 x

1 0 5367.606

2 1 6222.826

> t.test(RoomRent~IsNewYearEve, data=cities)

Welch Two Sample t-test

data: RoomRent by IsNewYearEve

t = -4.1793, df = 2065, p-value = 3.046e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1256.5297 -453.9099

sample estimates:

mean in group 0 mean in group 1

5367.606 6222.826

The mean room rent on New year eve is more than the mean room rent on other days.

P value is less than 0.05 so we can reject the null hypothesis and conclude that the room rent on new year eve is more than the room rent on other days.

3)

H0: The room rent for hotels having free breakfast and for hotels without free breakfast is the same.

H1: The room rent for hotels having free breakfast is more than the room rent for hotels without free breakfast

aggregate(cities$RoomRent, by=list(cities$FreeBreakfast), mean)

Group.1 x

1 0 5573.790

2 1 5420.044

> t.test(RoomRent~FreeBreakfast, data=cities)

Welch Two Sample t-test

data: RoomRent by FreeBreakfast

t = 0.98095, df = 6212.3, p-value = 0.3267

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-153.5017 460.9935

sample estimates:

mean in group 0 mean in group 1

5573.790 5420.044

The mean room rent for hotels with and without free breakfast is almost similar.

High p value(greater than 0.05) means we fail to reject the null hypothesis.

4)

H0: The room rent of hotels having free wifi and room rent of hotels without free wifi are almost similar

H1: The room rent of hotels having free wifi is greater than the room rent of hotels without free wifi

aggregate(cities$RoomRent, by=list(cities$FreeWifi), mean)

Group.1 x

1 0 5380.004

2 1 5481.518

> t.test(RoomRent~FreeWifi, data=cities)

Welch Two Sample t-test

data: RoomRent by FreeWifi

t = -0.76847, df = 1804.7, p-value = 0.4423

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-360.5977 157.5701

sample estimates:

mean in group 0 mean in group 1

5380.004 5481.518

The mean room rent of hotels with free wifi and hotels without free wifi are almost similar

We get a high p value(more than 0.05) so we fail to reject the null hypothesis.

5)

H0: The room rent of hotels in tourist cities and non tourist cities is the same

H1: The room rent of hotels in tourist cities is more than room rent of hotels in non tourist cities

aggregate(cities$RoomRent, by=list(cities$IsTouristDestination), mean)

Group.1 x

1 0 4111.003

2 1 6066.024

> t.test(RoomRent~IsTouristDestination, data=cities)

Welch Two Sample t-test

data: RoomRent by IsTouristDestination

t = -19.449, df = 12888, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2152.059 -1757.983

sample estimates:

mean in group 0 mean in group 1

4111.003 6066.024

The mean room rent of hotels in tourist cities is more than the mean room rent of hotels in non tourist cities

A very low p value(less than 0.05) helps us to reject null hypothesis and conclude that the room rent of hotels in tourist cities is more than the room rent of hotels in non tourist cities

6)

H0: The room rent of hotels in metro cities and non metro cities is the same

H1: The room rent of hotels in metro cities is more than the room rent of hotels in non metro cities

aggregate(cities$RoomRent, by=list(cities$IsMetroCity), mean)

Group.1 x

1 0 5782.794

2 1 4696.073

> t.test(RoomRent~IsMetroCity, data=cities)

Welch Two Sample t-test

data: RoomRent by IsMetroCity

t = 10.721, df = 13224, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

888.0308 1285.4102

sample estimates:

mean in group 0 mean in group 1

5782.794 4696.073

The mean room rent of hotels in metro cities is more than the mean room rent of hotels in non metro cities

A very low P value(less than 0.05) helps us to reject null hypothesis and conclude that the room rent of hotels in metro cities is more than the room rent of hotels in non metro cities

7)

H0: The room rent of hotels having swimming pool and without swimming pool is the same

H1: The room rent of hotels having swimming pool is more than the room rent of hotels without swimming pool

aggregate(cities$RoomRent, by=list(cities$HasSwimmingPool), mean)

Group.1 x

1 0 3775.566

2 1 8549.052

> t.test(RoomRent~HasSwimmingPool, data=cities)

Welch Two Sample t-test

data: RoomRent by HasSwimmingPool

t = -29.013, df = 5011.3, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-5096.030 -4450.942

sample estimates:

mean in group 0 mean in group 1

3775.566 8549.052

The mean room rent of hotels having swimming pool is more than the room rent of hotels without swimming pool

A very low P value(less than 0.05) helps us reject null hypothesis and conclude the room rent of hotels with swimming pool is more than the room rent of hotels without swimming pool.

***So we can conclude from the t tests that IsMetroCity, IsTouristDestination, IsNewYearEve and HasSwimmingPool have p values less than 0.05 and are statistically significant. Hence they play an important role in determining the Room rent of a particular hotel.***

**REGRESSION MODEL**

For the first model we consider most of the variables in the data set.

We ignore the cityrank because it is already covered in the city name, and date because it is covered in IsNewYearEve

We also exclude the Hotel name, address,pincode and description as they are specific to a particular hotel and cannot be interpreted as factors which affect the room rent.

**MODEL 1**

|  |
| --- |
| model1 <- RoomRent ~ CityName + Population + IsMetroCity + IsTouristDestination +  IsWeekend + IsNewYearEve + StarRating + Airport + FreeWifi + FreeBreakfast +  HotelCapacity + HasSwimmingPool  fit <- lm(model1, data=cities)  > summary(fit)  Call:  lm(formula = model1, data = cities)  Residuals:  Min 1Q Median 3Q Max  -11041 -2162 -461 1130 309155  Coefficients: (1 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) -10827.326 801.995 -13.500 < 2e-16 \*\*\*  CityNameAhmedabad 1100.123 743.874 1.479 0.13919  CityNameAmritsar 4.874 637.026 0.008 0.99390  CityNameBangalore -854.505 736.170 -1.161 0.24577  CityNameBhubaneswar 1317.537 897.394 1.468 0.14208  CityNameChandigarh 945.822 768.854 1.230 0.21865  CityNameChennai -1042.201 758.258 -1.374 0.16932  CityNameDarjeeling 1148.884 675.809 1.700 0.08915 .  CityNameDelhi -1409.914 1042.217 -1.353 0.17614  CityNameGangtok -3537.509 1107.919 -3.193 0.00141 \*\*  CityNameGoa 3510.467 422.252 8.314 < 2e-16 \*\*\*  CityNameGuwahati -352.962 994.427 -0.355 0.72264  CityNameHaridwar -756.626 989.963 -0.764 0.44470  CityNameHyderabad 232.698 734.909 0.317 0.75153  CityNameIndore 1293.113 846.862 1.527 0.12680  CityNameJaipur 2562.563 389.815 6.574 5.09e-11 \*\*\*  CityNameJaisalmer 3230.615 506.276 6.381 1.82e-10 \*\*\*  CityNameJodhpur 6390.012 533.060 11.987 < 2e-16 \*\*\*  CityNameKanpur 1464.437 1751.660 0.836 0.40315  CityNameKochi 1076.601 427.443 2.519 0.01179 \*  CityNameKolkata -179.538 911.530 -0.197 0.84386  CityNameLucknow 2535.484 883.188 2.871 0.00410 \*\*  CityNameMadurai 1361.387 693.063 1.964 0.04952 \*  CityNameManali 187.697 554.998 0.338 0.73522  CityNameMangalore 1294.042 925.345 1.398 0.16200  CityNameMumbai 859.057 1060.030 0.810 0.41772  CityNameMunnar 15.984 897.479 0.018 0.98579  CityNameMysore -815.862 1152.342 -0.708 0.47896  CityNameNainital 207.220 764.665 0.271 0.78640  CityNameOoty -763.217 955.658 -0.799 0.42452  CityNamePanchkula 159.308 1052.711 0.151 0.87972  CityNamePune 738.620 722.808 1.022 0.30686  CityNamePuri 161.111 1012.954 0.159 0.87363  CityNameRajkot 908.118 884.755 1.026 0.30472  CityNameRishikesh 696.702 774.429 0.900 0.36833  CityNameShimla 1461.467 544.626 2.683 0.00730 \*\*  CityNameSrinagar 2697.017 1084.719 2.486 0.01292 \*  CityNameSurat 519.774 989.455 0.525 0.59937  CityNameThiruvanthipuram 728.757 454.924 1.602 0.10919  CityNameThrissur -2827.231 1255.575 -2.252 0.02436 \*  CityNameUdaipur 5407.703 454.775 11.891 < 2e-16 \*\*\*  CityNameVaranasi 4719.234 518.990 9.093 < 2e-16 \*\*\*  Population NA NA NA NA  IsMetroCity 1112.495 978.811 1.137 0.25574  IsTouristDestination 776.672 600.488 1.293 0.19590  IsWeekend -97.516 121.832 -0.800 0.42349  IsNewYearEve 874.767 178.051 4.913 9.08e-07 \*\*\*  StarRating 3761.325 111.370 33.773 < 2e-16 \*\*\*  Airport 42.844 7.725 5.546 2.97e-08 \*\*\*  FreeWifi 729.163 224.852 3.243 0.00119 \*\*  FreeBreakfast -4.775 127.487 -0.037 0.97013  HotelCapacity -10.090 1.027 -9.824 < 2e-16 \*\*\*  HasSwimmingPool 1862.276 166.492 11.185 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 6463 on 13180 degrees of freedom  Multiple R-squared: 0.2263, Adjusted R-squared: 0.2233  F-statistic: 75.57 on 51 and 13180 DF, p-value: < 2.2e-16 |
| We notice that the variables such as CityName, Population, IsMetroCity, IsTouristDestination,  IsWeekend, FreeBreakfast are not statistically significant as indicated by the lack of stars on  the right hand side.  So while constructing the next model, we eliminate the above mentioned variables.  **MODEL 2**  model2 <- RoomRent ~ IsNewYearEve + StarRating + Airport + FreeWifi + HotelCapacity + HasSwimmingPool  > fit <- lm(model2, data=cities)  > summary(fit)  Call:  lm(formula = model2, data = cities)  Residuals:  Min 1Q Median 3Q Max  -10934 -2283 -879 986 310533  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -7574.167 394.531 -19.198 < 2e-16 \*\*\*  IsNewYearEve 835.593 175.987 4.748 2.08e-06 \*\*\*  StarRating 3516.463 111.595 31.511 < 2e-16 \*\*\*  Airport 25.588 2.601 9.837 < 2e-16 \*\*\*  FreeWifi 210.740 223.153 0.944 0.345  HotelCapacity -14.750 1.005 -14.676 < 2e-16 \*\*\*  HasSwimmingPool 2716.492 158.487 17.140 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 6681 on 13225 degrees of freedom  Multiple R-squared: 0.1703, Adjusted R-squared: 0.1699  F-statistic: 452.4 on 6 and 13225 DF, p-value: < 2.2e-16  We notice that FreeWifi is not statistically significant as indicated by the lack of  Stars. So while building the next model, we eleiminate the FreeWifi variable.  **MODEL 3**  model3 <- RoomRent ~ IsNewYearEve + StarRating + Airport + HotelCapacity + HasSwimmingPool  > fit <- lm(model3, data=cities)  > summary(fit)  Call:  lm(formula = model3, data = cities)  Residuals:  Min 1Q Median 3Q Max  -10920 -2281 -882 991 310542  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -7388.523 342.068 -21.600 < 2e-16 \*\*\*  IsNewYearEve 835.597 175.986 4.748 2.08e-06 \*\*\*  StarRating 3521.996 111.441 31.604 < 2e-16 \*\*\*  Airport 25.337 2.588 9.791 < 2e-16 \*\*\*  HotelCapacity -14.775 1.005 -14.706 < 2e-16 \*\*\*  HasSwimmingPool 2708.650 158.269 17.114 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 6681 on 13226 degrees of freedom  Multiple R-squared: 0.1702, Adjusted R-squared: 0.1699  F-statistic: 542.7 on 5 and 13226 DF, p-value: < 2.2e-16 |
| |  | | --- | |  | |

We find that model 3 is the best fit model, since all the independent variables are statistically significant.

Based on the regression analysis we conclude that the IsNewYearEve, StarRating, Airport, HotelCapacity, HasSwimmingPool are the most independent variables to determine the RoomRent of a particular hotel.

**Regression model equation**

**RoomRent = (835.597)IsNewYearEve + (3521.996)StarRating + (25.337)Airport + (-14.775) HotelCapacity + (2708.650) HasSwimmingPool + (-7388.523)**