AirBnB Data Analysis

Sara Petrov

# Research Question/Focus

### What variables are most directly linked with price?

#### The following variables were analyzed:

#### 1. Distance from city center

#### 2. Superhost Status

#### 3. Room type

#### 4. Cleanliness Rating

# Prepare, Clean and Process Data

### Uploaded libraries needed

library("ggplot2")

### Uploaded individual city AirBnB data, added a city column to each dataframe and joined them into one European City dataframe.

london1<-read.csv("london\_weekdays.csv")  
london2<-read.csv("london\_weekends.csv")  
London<-rbind(london1,london2)  
remove(london1,london2)  
London[1]<-NULL  
London$City<-"London"

Paris1<-read.csv("paris\_weekdays.csv")  
Paris2<-read.csv("paris\_weekends.csv")  
Paris<-rbind(Paris1,Paris2)  
remove(Paris1,Paris2)  
Paris[1]<-NULL  
Paris$City<-"Paris"

vienna1<-read.csv("vienna\_weekdays.csv")  
vienna2<-read.csv("vienna\_weekends.csv")  
Vienna<-rbind(vienna1,vienna2)  
remove(vienna1,vienna2)  
Vienna[1]<-NULL  
Vienna$City<-"Vienna"

barcelona1<-read.csv("barcelona\_weekdays.csv")  
barcelona2<-read.csv("barcelona\_weekends.csv")  
Barcelona<-rbind(barcelona1,barcelona2)  
remove(barcelona1,barcelona2)  
Barcelona[1]<-NULL  
Barcelona$City<-"Barcelona"

budapest1<-read.csv("budapest\_weekdays.csv")  
budapest2<-read.csv("budapest\_weekends.csv")  
Budapest<-rbind(budapest1,budapest2)  
remove(budapest1,budapest2)  
Budapest[1]<-NULL  
Budapest$City<-"Budapest"

lisbon1<-read.csv("lisbon\_weekdays.csv")  
lisbon2<-read.csv("lisbon\_weekends.csv")  
Lisbon<-rbind(lisbon1,lisbon2)  
remove(lisbon1,lisbon2)  
Lisbon[1]<-NULL  
Lisbon$City<-"Lisbon"

rome1<-read.csv("rome weekdays.csv")  
rome2<-read.csv("rome weekends.csv")  
Rome<-rbind(rome1,rome2)  
remove(rome1,rome2)  
Rome[1]<-NULL  
Rome$City<-"Rome"

athens1<-read.csv("athens\_weekdays.csv")  
athens2<-read.csv("athens\_weekends.csv")  
Athens<-rbind(athens1,athens2)  
remove(athens1,athens2)  
Athens[1]<-NULL  
Athens$City<-"Athens"

berlin1<-read.csv("berlin\_weekdays.csv")  
berlin2<-read.csv("berlin\_weekends.csv")  
Berlin<-rbind(berlin1,berlin2)  
remove(berlin1,berlin2)  
Berlin[1]<-NULL  
Berlin$City<-"Berlin"

amsterdam1<-read.csv("amsterdam\_weekdays.csv")  
amsterdam2<-read.csv("amsterdam\_weekends.csv")  
Amsterdam<-rbind(amsterdam1,amsterdam2)  
remove(amsterdam1,amsterdam2)  
Amsterdam[1]<-NULL  
Amsterdam$City<-"Amsterdam"

EuropeanCities<-do.call("rbind",list(Amsterdam,Athens,Barcelona,Berlin,Budapest,Lisbon,London,Paris,Rome,Vienna))

# Analysis of Data

## Analyze First Variable: Determine how distance from city center impacts price

range(EuropeanCities$dist)

## [1] 0.01504452 25.28455675

mean(EuropeanCities$dist)

## [1] 3.191285

sd(EuropeanCities$dist)

## [1] 2.393803

#### **Find mean price for different distances from city center and compare.**

mean(EuropeanCities[EuropeanCities$dist <=1,'realSum'])

## [1] 297.2791

mean(EuropeanCities[EuropeanCities$dist <=3,'realSum'])

## [1] 283.9003

mean(EuropeanCities[EuropeanCities$dist >=3,'realSum'])

## [1] 274.6435

mean(EuropeanCities[EuropeanCities$dist >=5,'realSum'])

## [1] 248.7961

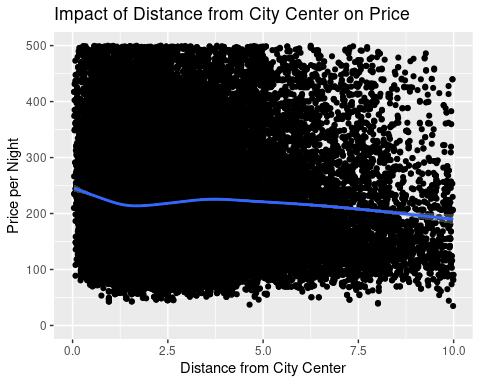
#### **Created scatterplot to test for negative correlation. Set limit on x and y axis to remove outliers.**

ggplot(EuropeanCities,aes(y=realSum, x=dist))+geom\_point()+stat\_smooth()+ylim(0,500)+xlim(0,10)+xlab("Distance from City Center")+ylab("Price per Night")+ggtitle("Impact of Distance from City Center on Price")

## `geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 6018 rows containing non-finite values (`stat\_smooth()`).

## Warning: Removed 6018 rows containing missing values (`geom\_point()`).



#### **Found P-Value and correlation coefficient**

ptest<-cor.test(EuropeanCities$realSum,EuropeanCities$dist, method ="pearson")  
ptest$p.value

## [1] 2.563113e-24

ptest$estimate

## cor   
## -0.04472541

### Summary of Results: Based on the comparison of means, the Pearson Correlation Coefficient, the p-value and the scatterplot, it is clear that the distance from the city center does not have a strong impact on the price per night. There is a slight negative correlation between the two variables based on the correlation coefficient but it is not significant.

## Analyze Second Variable: Superhost Status

mean(EuropeanCities[EuropeanCities$host\_is\_superhost =='False','realSum'])

## [1] 304.7729

mean(EuropeanCities[EuropeanCities$host\_is\_superhost =='True','realSum'])

## [1] 266.8417

#### **Made a copy of dataframe in order to change true and false variables to numeric in order to do a Pearson Correlation Coefficient test.**

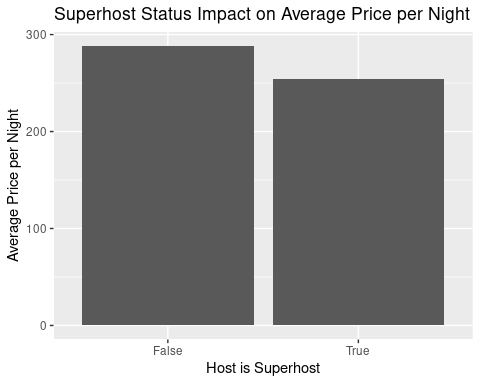
Copy<-data.frame(EuropeanCities)  
Copy$host\_is\_superhost[Copy$host\_is\_superhost=="True"]<-1  
Copy$host\_is\_superhost[Copy$host\_is\_superhost=="TRUE"]<-1  
Copy$host\_is\_superhost[Copy$host\_is\_superhost=="False"]<--1  
Copy$host\_is\_superhost[Copy$host\_is\_superhost=="FALSE"]<--1  
Copy$host\_is\_superhost<-as.numeric(Copy$host\_is\_superhost)

#### **Changed capital versions of false and true to lowercase**

Copy2<-data.frame(EuropeanCities)  
Copy2$host\_is\_superhost[Copy2$host\_is\_superhost=="FALSE"]<-"False"  
Copy2$host\_is\_superhost[Copy2$host\_is\_superhost=="TRUE"]<-"True"

#### **Created bar graph of average rooms price per night categorized by superhost status**

ggplot(Copy2, aes(x = factor(host\_is\_superhost), y = realSum)) +   
 stat\_summary(fun = "mean", geom = "bar")+ylab("Average Price per Night")+xlab("Host is Superhost")+ggtitle("Superhost Status Impact on Average Price per Night")



#### **Found P-Value and Correlation Coefficient**

ptesthost<-cor.test(Copy$realSum,Copy$host\_is\_superhost, method ="pearson")  
ptesthost$p.value

## [1] 4.345181e-25

ptesthost$estimate

## cor   
## -0.04547789

### Summary of Results: Based on analysis, being a superhost does not have a significant impact on price per night.

### Analyze Third variable: Impact of Room Type on Price Per Night

mean(EuropeanCities[EuropeanCities$room\_type =='Private room','realSum'])

## [1] 204.9402

mean(EuropeanCities[EuropeanCities$room\_type =='Shared room','realSum'])

## [1] 143.5662

mean(EuropeanCities[EuropeanCities$room\_type =='Entire home/apt','realSum'])

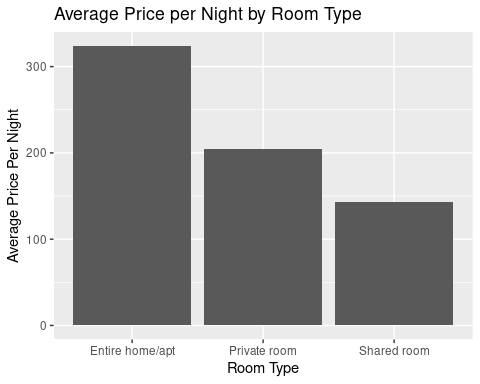
## [1] 324.3152

#### **Made a copy of dataframe in order to change room type variables to numeric in order to do a Pearson Correlation Coefficient test.**

CopyRT<-data.frame(EuropeanCities)  
CopyRT$room\_type[CopyRT$room\_type=="Private room"]<-1  
CopyRT$room\_type[CopyRT$room\_type=="Shared room"]<-2  
CopyRT$room\_type[CopyRT$room\_type=="Entire home/apt"]<-3  
CopyRT$room\_type<-as.numeric(CopyRT$room\_type)

#### **Created bar graph of average rooms price per night categorized by room type**

ggplot(EuropeanCities, aes(x = factor(room\_type), y = realSum)) +   
 stat\_summary(fun = "mean", geom = "bar")+ylab("Average Price Per Night")+xlab("Room Type")+ggtitle("Average Price per Night by Room Type")



#### **Found P-Value and Correlation Coefficient**

ptestroom<-cor.test(CopyRT$realSum,CopyRT$room\_type, method ="pearson")  
ptestroom$p.value

## [1] 0

ptestroom$estimate

## cor   
## 0.1753155

### Summary of Results:From analysis, it appears that entire home/apt the price is the most expensive. It is twice as much as a shared room. In addition, the the correlation is strong based on the p value of 0 and correlation coefficient of .18. It can be said that room type has a strong positive correlation with price per night.

### Analyze Fourth variable: Impact of Cleanliness Rating on Price Per Night

#### **Found mean price for different cleanliness rating ranges.**

mean(EuropeanCities[EuropeanCities$cleanliness\_rating <=2,'realSum'])

## [1] 338.6613

mean(EuropeanCities[EuropeanCities$cleanliness\_rating <=5,'realSum'])

## [1] 302.7132

mean(EuropeanCities[EuropeanCities$cleanliness\_rating >=5,'realSum'])

## [1] 279.5942

mean(EuropeanCities[EuropeanCities$cleanliness\_rating >=8,'realSum'])

## [1] 278.908

#### **Found p-value and correlation coefficient**

ptestclean<-cor.test(EuropeanCities$realSum,EuropeanCities$cleanliness\_rating, method ="pearson")  
ptestclean$p.value

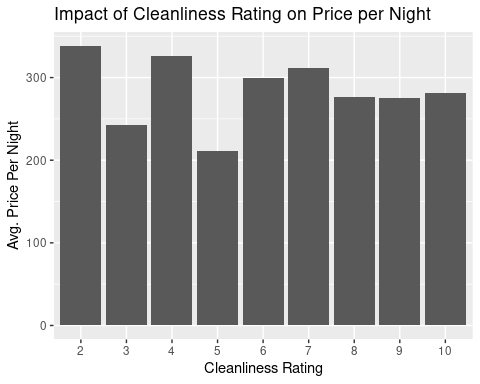
## [1] 0.1666841

ptestclean$estimate

## cor   
## -0.00608183

#### **created scatterplot. Set limit on y axis to remove outliers.**

ggplot(EuropeanCities, aes(x = factor(cleanliness\_rating), y = realSum)) +   
 stat\_summary(fun = "mean", geom = "bar")+ylab("Avg. Price Per Night")+xlab("Cleanliness Rating")+ggtitle("Impact of Cleanliness Rating on Price per Night")



### Summary of Results: As can be seen on the bar graph, there appears to be know significant difference in room price based on cleanliness rating. In addition, the p-value and correlation coefficient show that there is a not a strong correlation between price per night and cleanliness rating.