

Assignment

Data Analytics Using SPSS

Objectives of this assignment

- Review basic steps in preparing data for statistical analysis
 - Create dummy variables (for categorical and continuous variables)
 - Aggregate data by groups using sum, mean, first or unique values
 - Calculate and interpret descriptive statistics (minimum, maximum, mean, mode, median, standard deviation, variance)
 - Create and interpret graphical illustrations (histogram, scatterplot)
- Apply linear regression analysis
 - Estimate linear regression model
 - Interpret regression coefficients and fit measures
 - Plotting and reporting estimates

Software

SPSS

Dataset

Booking.com data (Please see the attachment)

Introduction

You are doing an internship with an online travel agency based in the Netherlands and you are being assigned to a team that is in the process of designing a new travel package that allows couples to spend a 1-night short getaway in Amsterdam. Given that last-minute travel is the up and coming trend for couple travelers, to accommodate this spontaneity, a key feature of this new travel package is that it provides the option for couples to make reservations for the package on a short notice. After spending several weeks brainstorming and researching on the different romantic last-minute activities that can be included in the travel package, the team is left with the accommodation options that will determine the notice period required for booking the travel package. Being the newbie (and also an intern) for the team, your task is to provide insights on an appropriate notice period based on a dataset that the team had gathered. Specifically, the team is interested in understanding the dynamics of the accommodation market in Amsterdam and factors that influence the availability of properties across the various accommodation options.

The dataset is at the notice period accommodation level, where 1 August 2018 is the date at which the search was conducted, and the accommodation options reflect properties in Amsterdam that are listed on Booking.com as of the search date.

For more details on the data, please refer to **Booking.com Data Dictionary 2020**. (Please see the attachment)

Some general tips for Assignment:

- Use SPSS syntax whenever possible! It is easier, less prone to error and makes data manipulation and analysis more efficient. You can get a good idea on how to use SPSS syntax in the solutions for the practice questions.
- A term you should know to decipher some of the comments in the SPSS syntax file, i.e., GUI = Graphic User Interface. For those of you who feel more comfortable using the graphic user interface, I have included some accompanying videos and “directions” to illustrate how you should go about clicking on the buttons in the interface for some syntax commands. These “directions” are denoted by the “GUI:...” in the SPSS syntax file.
- To aggregate data by groups, first include the grouping variable in the “Break Variable(s)” button in SPSS. Then depending on the context of the question, select the aggregate function (i.e., by clicking on the “Function” button) to aggregate data using sum and mean values
- Always think about the aggregation level of the dataset and compare it to the aggregation level that the question requires!
- Don’t go crazy with decimal points! Providing your responses rounded off to 3 decimal points (i.e., 3 d. p.) is acceptable for this assignment (i.e., if applicable). However, please also take note that in certain circumstances, it will be more appropriate for you to round it off to a whole number instead. Also, don’t forget the units (e.g., km etc.) in your responses!
- In this assignment, we will use the significance level of 5% to evaluate the statistical significance of test results.

Sample Practice Questions

To provide you with some guidance for Assignment, below are some sample practice questions that will give you a sense of what to expect for the actual assignment. Consistent with the actual assignment, the practice questions can be classified into two broad categories – descriptive statistics and regression analysis. While the purpose of examining descriptive statistics is to obtain a more in-depth understanding of the characteristics of the dataset, regression analysis serves to examine the relationships between one or more of the variables within the dataset and in restricted circumstances, to infer causal relationships amongst these variables. (**tip** : You should read through **Appendix A: Dataset** and/or the **Booking.com Data Dictionary**) to get an in-depth understanding of the dataset before attempting the sample practice questions.)

The solutions to the sample practice questions are provided in **Appendix B: Solutions to Sample Practice Questions** of this document.

Descriptive Statistics

1. Original aggregation level: Notice period accommodation level
 - a. How many unique accommodation options are there in total? (**tip** : You should consider all notice periods.)
 - b. Which is the notice period with the least amount of unique accommodation options left? (**tip** : Remember to remove accommodation options with no rooms left! Also, note that rooms and accommodation options have different meanings.)
 - c. Amongst the accommodation options that are available for the notice period that you have identified in 1b, what is the average number of people who are also looking at the accommodation at the time of the search? (**tip** : See tip for 1b.)
 - d. Relatedly, amongst the accommodation options that are available for the notice period that you have identified in 1b, what is the variation in the number of people who are also looking at the accommodation at the time of the search from the mean? (**tip** : Think about the definitions of mean, standard deviation and variance. Also, see tip for 1b.)
2. Derived aggregation level: Notice period (**tip** : Depending on the context of the question, you would need to aggregate the data to the notice period level either by taking the sum or by average values.)
 - a. Create and interpret the scatterplot of the number of accommodation options with free cancellation for the cheapest room available per notice period; are the results of the scatterplot within your expectations?

- b. Create the scatterplot of the review ratings of accommodation options per notice period; what can you conclude from the scatterplot? Please provide your interpretation of the scatterplot and discuss any managerial insights/implications that you can draw from your interpretation of the scatterplot.
3. Derived aggregation level: Accommodation level (**tip** : Given the context of the following questions, you should aggregate the data to the accommodation level by the average values to avoid double counting.)
 - a. Which accommodation option(s) has the least number of reviews and the lowest review rating? Please explain how you arrive at your answer. (**tip** : An accommodation with no reviews will naturally not have any review ratings. As such, for this question, we are not interested in these accommodation options.)
 - b. How many accommodation options are located in Oud Zuid?
 - c. Create the histogram of the review ratings of accommodation options by star rating; what can you conclude from the histogram? (**tip** : Note that star ratings are not applicable for apartment-type accommodation options.)
 - d. What is the mean, minimum and maximum distance to town for accommodation options located in Oud-West? (**tip** : What is the unit for measuring distance to town? If you are unsure, look at Appendix A and/or the data dictionary again and provide this unit in your response.)

Regression Analysis

4. Estimation

Suppose your team manager is interested to find out more about the characteristics that influence the availability of rooms. Use linear regression analysis and estimate the following model:

$$\begin{aligned}
 \text{roomsleft}_{it} = & \beta_0 + \beta_1 \text{threedaysnp}_{it} + \beta_2 \text{oneweeknp}_{it} + \beta_3 \text{twoweeksn}_{it} \\
 & + \beta_4 \text{onemonthnp}_{it} + \beta_5 \text{threemonthsn}_{it} + \beta_6 \text{sixmonthsn}_{it} + \beta_7 \text{lownr}_i \\
 & + \beta_8 \text{highnr}_i + \beta_9 \text{peoplelooking}_{it} + \beta_{10} \text{highdemand}_{it} + \beta_{11} \text{metro}_i \\
 & + \beta_{12} \text{town}_i + \beta_{13} \text{freecancel}_{it} + \beta_{14} \text{apartment}_i + \beta_{15} \text{greatfortwo}_i + \epsilon_{it}
 \end{aligned}$$

where roomsleft_{it} is the dependent variable and indicates the number of rooms left for accommodation i at time t , threedaysnp_{it} represents a 3-days' notice period for accommodation i at time t , oneweeknp_{it} represents a 1-week notice period, twoweeksn_{it} represents a 2-weeks' notice period, onemonthnp_{it} represents a 1-month notice period, sixmonthsn_{it} represents a 6-months' notice period, lownr_i suggests that there is a lower number of reviews for accommodation i , highnr_i suggests that there is a higher number of reviews, $\text{peoplelooking}_{it}$ indicates the number of people that are looking at the accommodation i at time t , highdemand_{it} suggests that the accommodation i is in high demand at time t , metro_i suggests that the accommodation i is close to a metro station, town_i indicates the distance of accommodation i to town, freecancel_{it} suggests that guest have the option to make a risk-free reservation of the cheapest room available for the accommodation i at time t ,

apartment_i suggests that the accommodation i is an apartment, greatfortwo_i suggests that the accommodation i is recommended for two travelers, and ϵ_{it} represents the error term for the model.

(tip : To answer this question, you will first need to prepare the data by creating additional variables, in particular, dummy variables. First, create notice period dummies to represent each notice period. Based on the model specified (above), the reference point is the 1-year notice period. Next, you will also have to create dummy variables to represent the high, low, medium levels of the number of reviews, with the reference point being the medium level, i.e., based on model specified above).

- a. Interpret your findings from the model results. (**tip :** Be sure to discuss the following: 1) Overall Model Significance, 2) Overall Model Fit, and 3) For each of the regression coefficients, evaluate the statistical significance, sign, and magnitude.)
- b. Based on your interpretation of the results from the model, what should your team manager do to ensure that there is a greater availability of rooms?
- c. Plot the coefficients of the notice period dummies in a bar chart.

Appendix A: Dataset

File format: .SAV (SPSS dataset)

The dataset consists of 7,010 observations comprising of accommodation options in Amsterdam that are listed on Booking.com on the search date, i.e., 1 August 2018. The 3 tables below denote the format as well as the description of the variables featured in the dataset. For a more detailed description of each of the variables (i.e., description with visuals), please refer to the data dictionary **Booking.com Data Dictionary**).

IDENTIFICATION VARIABLES		
Variable Name	Format	Description
np	Numeric	Notice period, measured by the number of days between the search date (i.e., 1 August 2018) and the check-in date, where 3 = search date is 3 days before the check-in date, 7 = search date is 1 week before the check-in date, 14 = search date is 2 weeks before the check-in date, 31 = search date is 1 month before the check-in date, 92 = search date is 3 months before the check-in date, 184 = search date is half a year before the check-in date, 365 = search date is a year before the check-in date.
accom	Character	Name of the accommodation.
add	Character	Address of the accommodation.
apartment	Numeric	Type of accommodation, where 1 = Apartment-type accommodation, 0 = Otherwise
town	Numeric	Distance to town from the accommodation (in km).
metro	Numeric	Proximity to metro, where 1 = accommodation is close to a metro station, 0 = otherwise.
starrating	Numeric	Star ratings of the accommodation. Note that the star ratings are not applicable to apartment-type accommodations.
prefer	Numeric	Preferred partner property, where 1 = Accommodation is a “Preferred Partner”, 0 = otherwise.
prom	Numeric	Promoted property, where 1 = Accommodation is a promoted property, 0 = otherwise.

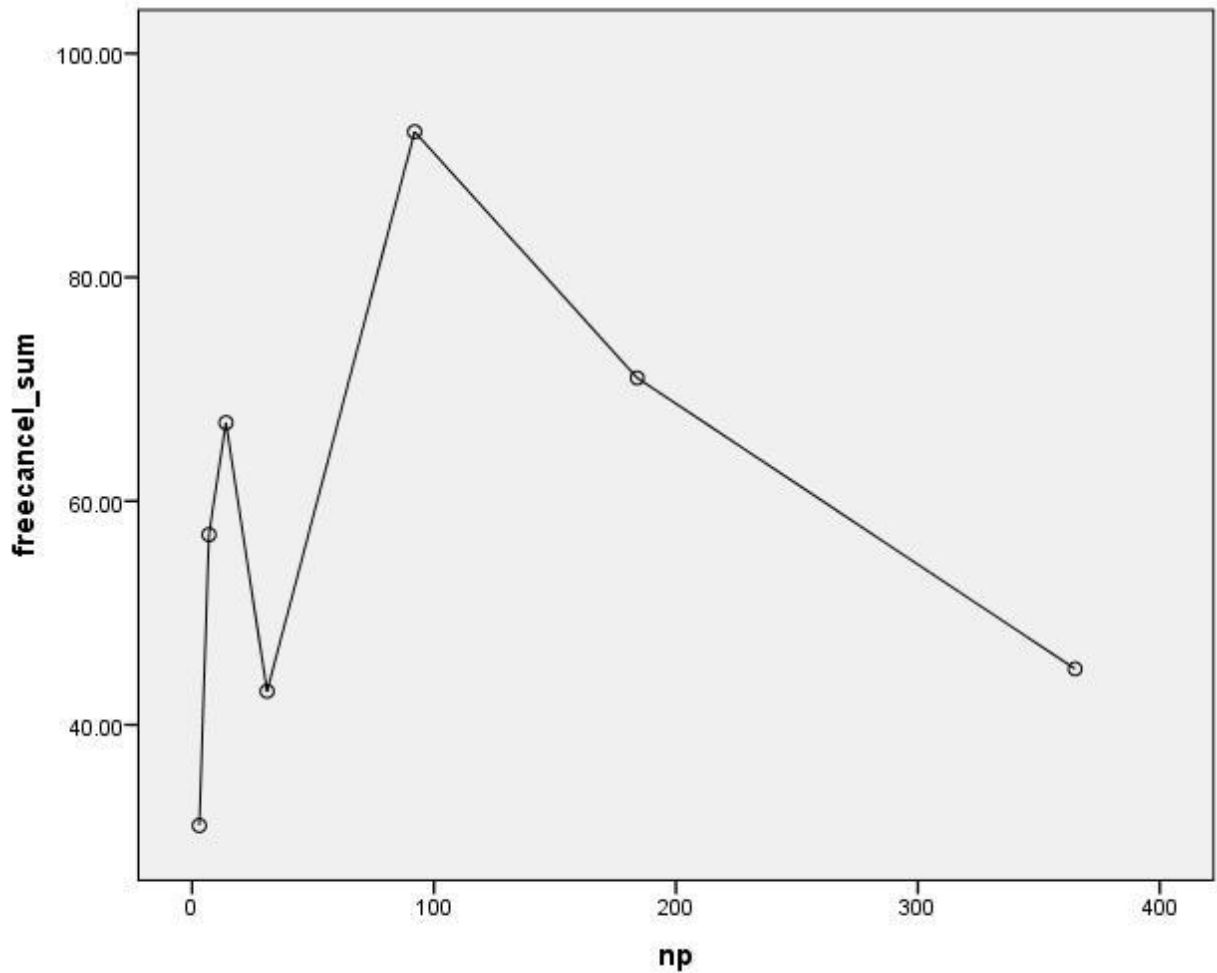
CONSUMER PREFERENCE		
Variable Name	Format	Description
bestseller	Numeric	Bestseller property, where 1 = Accommodation is a bestseller, 0 = Otherwise
peoplelooking	Numeric	Number of people also looking the accommodation at the time of search
highdemand	Numeric	High demand property, where 1 = Accommodation is in high demand, 0 = Otherwise
roomsleft	Numeric	Number of rooms left
numreviews	Numeric	Number of reviews
reviewrating	Numeric	Review rating
reviewcat	Character	Review category
locrating	Numeric	Location rating
greatfortwo	Numeric	Recommended for two travelers, where 1 = Accommodation is a recommended property for two travelers, 0 = Otherwise
guestfav	Numeric	Guest favorite, where 1 = Accommodation is a guest favorite, 0 = Otherwise
percentme	Numeric	Percentage of guest reviewers that had their expectations of this accommodation met or exceeded
DETAILS ON THE CHEAPEST ROOM		
Variable Name	Format	Description
price	Numeric	Price of the cheapest room available in Euros (€)
pprice	Numeric	Previous price of the cheapest room available in Euros (€)
roomtype	Numeric	Room type of the cheapest room available
freecancel	Numeric	Free cancellation, where 1 = Guests have the option to make a risk-free reservation of the cheapest room available, i.e., free cancellation is possible, 0 = Otherwise
greatvalue	Numeric	Great Value Today, where 1 = Accommodation is of great value on the selected check-in dates, 0 = Otherwise
roombanner	Character	Room banner (also acts as the button to navigate to the rooms available) that displays certain messages
roomdes1	Character	First description of the details of the cheapest room available
roomdes2	Character	Second description of the details of the cheapest room available

Appendix B: Solutions to Sample Practice Questions

Descriptive Statistics

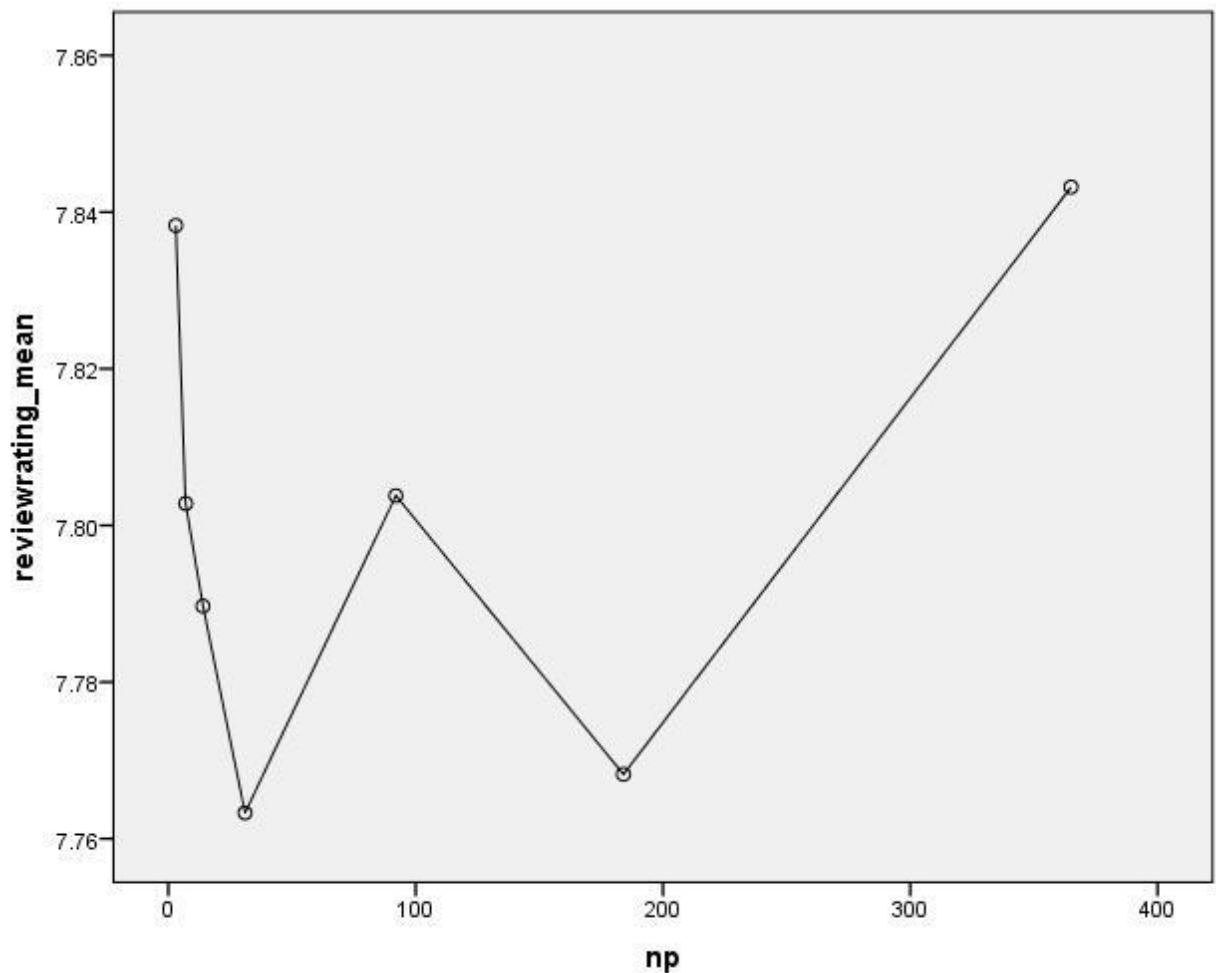
- 1a. 1,305 accommodation options
- 1b. Notice period = 365, i.e., 1 year before the check-in date. There are only 65 unique accommodation options available if the search is conducted using this notice period.
- 1c. The average number of people also looking at the accommodation at the time of the search for the 1-year notice period (i.e., = 365) is 0.785. Since we are talking about the number of **people** in this question, you should provide the answer rounded off to the nearest whole number. Therefore, the correct answer for this question is 1 person (**tip** : Note that this logic applies to your responses in the actual assignment as well).
- 1d. The variation of the number of people also looking at the accommodation at the time of the search for the 1-year notice period (i.e., = 365) from the mean implies that you need to obtain its variance which is 5.297.

- 2a. To obtain the scatterplot of the number of accommodation options with free cancellation for the cheapest room available (y-axis) by the notice period (x-axis), you will need to aggregate the data to the notice period level by taking the sum of the number of accommodation options with free cancellation for the cheapest room available (i.e., it is a dummy variable consisting of only 1 and 0, hence adding the ones will give you the total number of accommodation options with free cancellation for the cheapest room available) per notice period. The scatterplot will be as follows:



With regards to the interpretation of the scatterplot, we can see that the peak in the number of accommodation options with free cancellation for the cheapest room available occurs during the 1-month notice period and the dip occurs during the 3-days' notice period. The point at which the dip occurs is consistent with expectations as it is understandable that it is not in the best interests of the accommodation to offer a risk-free option at a very short notice. Following this logic, the point at which the peak occurs is not intuitive as one might expect the peak to occur perhaps in the 1-year or even the 6-months' notice period.

2b. To obtain the scatterplot of the review ratings of accommodation options (y-axis) by the notice period (x-axis), you will need to aggregate the data to the notice period level by taking the average of the review ratings of accommodation options (i.e., taking the sum of review ratings in this context do not make sense as the review rating of an accommodation can range from 1 to 10 and anything above 10 is not interpretable) per notice period. The scatterplot will be as follows:



In interpreting the scatterplot, we can observe that the peak in the review ratings of accommodation options occurs in the 1-year notice period and the dip occurs in the 1-month notice period. Assuming review ratings reflect the quality of the accommodation options, the scatterplot suggests that there is a greater number of higher quality accommodation options available at the 1-year notice period. As such, in terms of the availability of higher quality accommodation options, a 1-year notice period will be the most ideal for the new travel package.

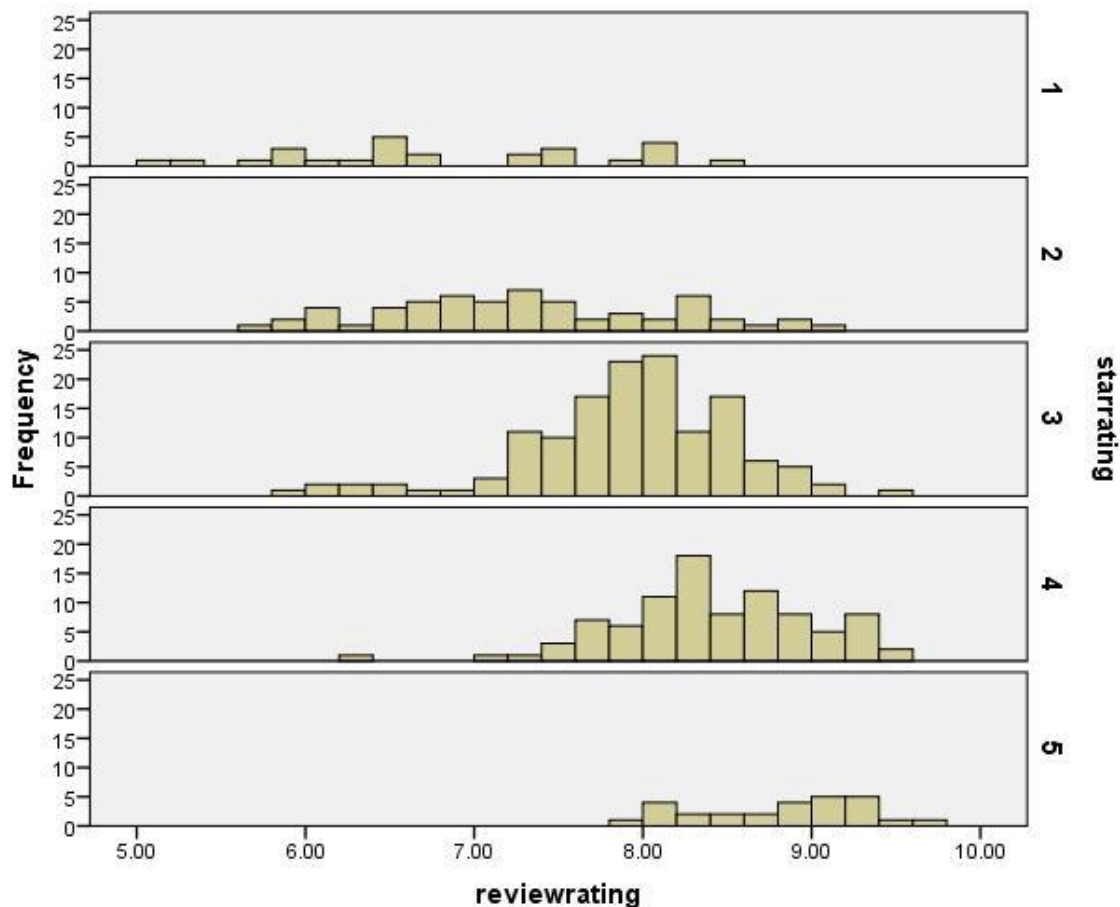
3a. Note that there are two solutions to this question.

If you first look for accommodation options with the lowest number of reviews (excluding zero), and then obtain the solution by examining the review ratings of these accommodation options, your response will be: Romantic garden studio near Westerpark and Jordaan. This is because the lowest non-zero value for number of reviews is 5 and within the accommodation options that only have 5 reviews, the lowest review rating is 6.5 which belongs to Romantic garden studio near Westerpark and Jordaan.

If you first look for accommodation options with the lowest review ratings (excluding zero), and then obtain the solution by examining the number of reviews of these accommodation options, your response will be: Cosy Studio. This is because the lowest non-zero value for review ratings is 4.6 and there is only 1 accommodation with this review rating and the number of reviews of Cosy Studio is 42.

3b. 166 accommodation options

3c. Before you create the histogram, you should remove observations with review ratings that are equivalent to zero, as these observations do not have any reviews and as such are not of our interest in the context of this question (**tip** : Note that including these observations is likely to create “noise” in the histograms as they are likely to skew the distributions). The histogram of the review ratings of accommodation options by star rating will be as follows:



While the distribution of review ratings for 1-star to 2-star hotels are skewed to the left, the distribution of review ratings for 3-star to 5-star hotels are skewed to the right. This implies that review ratings tend to be lower for 1-star to 2-star hotels but higher for 3-star to 5-star hotels. We can also observe that the number of 1-star and 5-star hotels are relatively lower compared to other accommodation options (**tip** : Look at the y-axis, i.e., “Frequency”).

3d. The SPSS syntax provided to you should generate this table:

Report			
town_first			
add_first	Mean	Minimum	Maximum
	9.2000	3.40	11.00
Amsterdam City Centre	6.3767	.10	11.00
Amsterdam Noord	7.5000	1.30	11.00
Bos en Lommer	6.9786	2.20	11.00
De Baarsjes	7.9375	2.20	11.00
Goudenveld-Slotervaart	7.7273	4.00	11.00
Oost	6.7703	2.00	11.00
Osdorp	9.0000	8.00	11.00
Oud Zuid	7.1000	1.40	11.00
Oud-West	6.6828	1.30	11.00
Sloterdijk	7.2500	3.50	11.00
Slotervaart	6.3000	3.40	11.00
Westerpark	7.1935	1.30	11.00
Westpoort	4.5462	3.50	11.00
Zeeburg	8.1619	1.60	11.00
Zuideramstel	6.1630	2.90	11.00
Zuidoost	8.6087	7.00	11.00
Total	6.7536	.10	11.00

Based on the table above, the mean distance to town for accommodation options located in Oud-West is 6.683 km, the maximum is 1.300 km and the minimum is 11.000 km. (**tip** : Note that the unit for measuring distance to town is *km.*)

Regression Analysis

4a. As stated in the tip for this question, the interpretation should be executed as follows.

Overall Model Significance:

ANOVA ^a					
Model		Sum of Squares	df	Mean Square	F
1	Regression	4203.037	15	280.202	209.597
	Residual	9350.039	6994	1.337	
	Total	13553.076	7009		

a. Dependent Variable: roomsleft

b. Predictors: (Constant), greatfortwo, twoweeksnr, metro, highdemand, apartment, freecancel, oneweeknr, onemonthnr, sixmonthsnr, peoplelooking, lownr, threedaysnr, town, threemonthsnr, highnr

The overall model significance assesses the extent to which the independent variables that are included in the model have any joint explanatory power. We can draw conclusions on the overall model significance by evaluating the statistical significance of the F -test (i.e., highlighted in yellow). As observed, the overall model is statistically significant.

Overall Model Fit:

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.557 ^a	.310	.309	1.156

a. Predictors: (Constant), greatfortwo, twoweeksnr, metro, highdemand, apartment, freecancel, oneweeknr, onemonthnr, sixmonthsnr, peoplelooking, lownr, threedaysnr, town, threemonthsnr, highnr

To draw conclusions about the overall model fit, we need to evaluate the R^2 (i.e., highlighted in yellow), which indicates the proportion of variance in the dependent variable (i.e., roomsleft) that is explained by the independent variables (i.e., the notice period and number of reviews dummies). The R^2 varies between 0 and 1, where 0 suggests that 0% of the variance is explained and 1 suggests that 100% of the variance is explained. As observed, the R^2 value of the model is 31% (or 0.310), indicating a slightly lower fit. In other words, the model explains only 31% of the variation in the number of rooms left. As such, results should be interpreted with caution especially when used to predict future values.

Interpretation for each of the regression coefficients:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.316	.056		5.670	.000
	threedaysnp	.213	.052	.054	4.110	.000
	oneweeknp	.430	.052	.108	8.243	.000
	twoweeksnp	.575	.052	.145	11.089	.000
	onemonthnp	.169	.052	.042	3.254	.001
	threemonthsnp	.621	.052	.156	11.923	.000
	sixmonthsnp	.332	.052	.084	6.422	.000
	lownr	-.014	.036	-.005	-.405	.685
	highnr	.650	.044	.220	14.634	.000
	peoplelooking	.160	.010	.172	15.312	.000
	highdemand	-1.606	.066	-.257	-24.458	.000
	metro	.166	.035	.049	4.773	.000
	town	-.054	.004	-.182	-12.947	.000
	freecancel	.882	.062	.148	14.268	.000
	apartment	-.077	.036	-.025	-2.139	.032
	greatertwo	-.117	.067	-.018	-1.749	.080

a. Dependent Variable: roomleft

The interpretation of the individual regression coefficients should be executed as follows. First, assess the statistical significance of a regression coefficient by evaluating the result of the -test (i.e., highlighted in purple). Second, evaluate the sign of the regression coefficient (i.e., whether it is a positive or a negative effect) and the magnitude of the regression coefficient (i.e., assess the size) by examining its unstandardized coefficient (i.e., highlighted in yellow). Note that if the regression coefficient is not statistically significant, there is no need to proceed to step 2. As such, the interpretation of the results of the regression coefficients are as follows:

- Notice Period Dummy Variables: Overall, all the notice period dummy variables have significantly higher number of rooms left when compared to the 1-year notice period, i.e., the reference notice period. The magnitude of their effects varies between 0.169 (1-month notice period) and 0.621 (3-months' notice period). In other words, while a 1-month notice period has 0.169 higher number of rooms left when compared to the reference notice period of 1 year, a 3-months' notice period has 0.621 higher number of rooms left when compared to the reference notice period of 1 year.

- **Dummy Variables for Number of Reviews:** While accommodation options with lower number of reviews do not significantly result in higher number of rooms left when compared with accommodation options with medium number of reviews, accommodation options with higher number of reviews have significantly higher number of rooms left, i.e., of magnitude 0.650, when compared with accommodation options with medium number of reviews.
- **People Looking and High Demand:** The number of people looking at the accommodation option has a positive and significant effect on the number of rooms left. In particular, as the number of people looking at the accommodation option goes up by 1, the number of rooms left goes up by 0.160. In contrast, accommodation options in high demand have significantly lower number of rooms left, i.e., of magnitude 1.605, as compared to those that are not in high demand.
- **Metro and Town:** While an accommodation option that is located near a metro station have significantly higher number of rooms left, i.e., 0.166, then one that is not, the proximity of an accommodation option to town has a negative and significant effect on the number of rooms left. Specifically, as the proximity of an accommodation option to town increases by 1 km, the number of rooms left goes down by 0.054.
- **Free Cancellation:** Accommodation options that provide guests the option to make a risk-free reservation of the cheapest room available have a significantly higher number of rooms left, i.e., of magnitude 0.882, when compared to those that do not.
- **Apartment:** We can also observe that apartment-type accommodation options have a significantly lower number of rooms left, i.e., of magnitude 0.077 when compared to non-apartment-type accommodation options, i.e., like hotels.
- **Great for Two:** Results suggest that whether an accommodation is recommended for two travelers does not influence the number of rooms left for that accommodation.

4b. To ensure that there is a greater availability of rooms, you can recommend your team manager to begin searching for non-apartment-type accommodation options that have higher number of reviews, greater number of people looking, not in high demand, close to a metro but not to town, and provides the option to make a risk-free reservation on the cheapest room available, 3 months before the desired check-in date.

4c. The plot of the notice period coefficients are as follows:

