Coding 2 Final Assignment

Introduction

This project looks at the website HostelWorld.com and involved scraping data off this website. Before going into what was scraped, a little background information on HostelWorld.com. This is a platform where individuals can find affordable accommodation while traveling and is particularly popular amongst students. Hostels offer dorm rooms and private rooms. The dorm rooms are always cheaper and will usually be a small bed in a room with 8 other beds. So, you reserve a bed in a dorm room. The private rooms are more expensive and are private rooms in the same accommodation location.

Hostel World collects a lot of ratings from their customers, which includes very specific rating information on different aspects of the Hostel or the city the Hostel is located in. This data can be very helpful in determining what may be associated with higher rates of books or higher prices. This project will focus on the change in prices over time, and what may be associated with the change in prices between hostels, cities, and time.

Web Scraping Process

For this project, Selenium was used for web scraping. Initially, other types of methods were attempted, such as scrapethat; however, these ideas were dropped since this project needed a large amount of data, and pages on Hostel World do not have a lot of data. Scrolling through pages is essential in certain regards. Using selenium allowed for easier access to opening a browser and processing through many listings across many different dates. Furthermore, using selenium is a very interesting method of scraping, and the researcher preferred utilizing this tool for experimentation purposes.

Three different web scraping functions were created for collecting the data. The first function, named scrape_hostel_data, takes inputs for the city, region, country, city ID, from date, to date, and number of guests. This is because these are found in the https link. The function sets an f-string with the link to a variable for easy of looping a list through this function. Overall, this function collects information such as the hostel name, overall rating, total rating count, private room price, dorm price, descriptive rating (good, very good, fabulous, superb), accommodation type (hostel, hotel, bed and breakfast), distance from the city center, city, country, and date. The limitation to this function is that the execution takes approximately 6 hours to run and collect the data. In the end, it was able to collect 77,000 listings. Each city was looked at for each day in January, April, July, and October, and price data was collected for each of these days for each individual hostel. The following is a list of the cities that were scraped:

> Berlin, Rome, Madrid, London, Amsterdam, Prague, Vienna, Budapest, Athens, Istanbul, Dublin, Brussels, Lisbon, Warsaw, Oslo, Stockholm, Helsinki, Copenhagen,

Riga, Tallinn, Vilnius, Tokyo, Seoul, Beijing, Shanghai, Bangkok, New Delhi, Mumbai, Kuala Lumpur, Singapore, & Dubai

Once this data was collected and the data was analyzed, there were some issues. One of the primary issues was that the links did not match the hostel names, and they were out of order. The objective was to use these links to collect information about each individual hostel. Each hostel has amenity scores and general ratings for different aspects of the hostel on a scale of 1-10, such as: security, location, staff, atmosphere, cleanliness, value for money, and facilities. Furthermore, there are location descriptors, staff descriptors, and cleanliness descriptors. These seemed interesting for data collection. As the previous function was not able to scrape details from the physical hostel pages due to an issue with the link collection, a new function was created so that the process did not need to be repeated for collecting price data (another limitation to the previous function). This new function is named scrape_hostel_links, and it takes continent, country, and city as input and accesses a different part of the Hostel World website that supplies data about each hostel in a given city. This function returns these scraped links, as well as the individual and detailed ratings for each hostel.

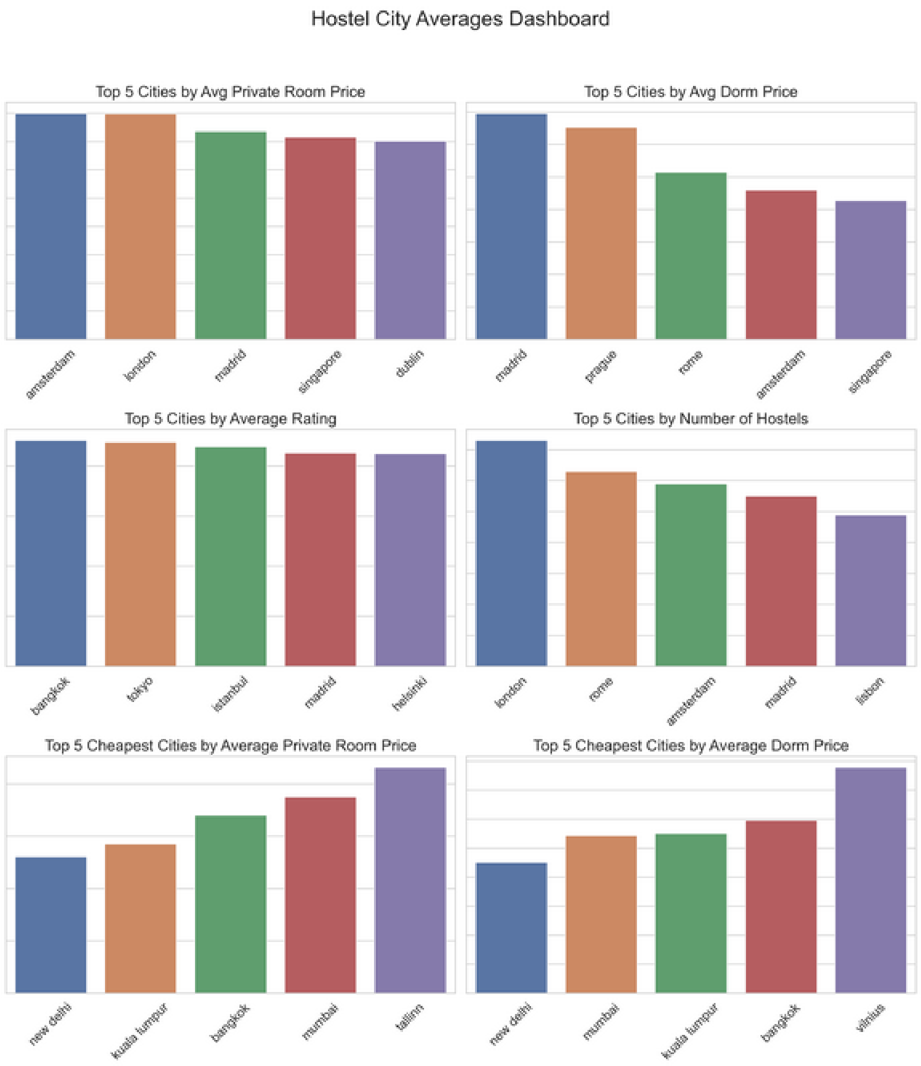
Finally, the Hostel World website offers ratings for how customers felt about the cities they stayed in. These ratings include a 1-10 scale for different aspects of a city: activities, eating out, shopping, chilling out, transport, culture, nightlife, and value for money. The function named scrape_city_details returns these detailed ratings for each given city.

Descriptive Visualizations for Aggregations

The data was aggregated by city to identify the top cities for specific categories. The following categories were visualized, along with the findings:

- *Top 5 Cities by Average Private Room Price*
  - Amsterdam, London, Madrid, Singapore, Dublin
- *Top 5 Cities by Average Dorm Price*
  - Madrid, Prague, Rome, Amsterdam, Singapore
- *Top 5 Cities by Average Rating*
  - Bangkok, Tokyo, Istanbul, Madrid, Helsinki
- *Top 5 Cities by Number of Hostels*
  - London, Rome, Amsterdam, Madrid, Lisbon
- *Top 5 Cheapest Cities by Average Private Room Price*
  - New Delhi, Kuala Lumpur, Bangkok, Mumbai, Tallinn
- *Top 5 Cheapest Cities by Average Dorm Price*
  - New Delhi, Mumbai, Kuala Lumpur, Bangkok, Vilnius

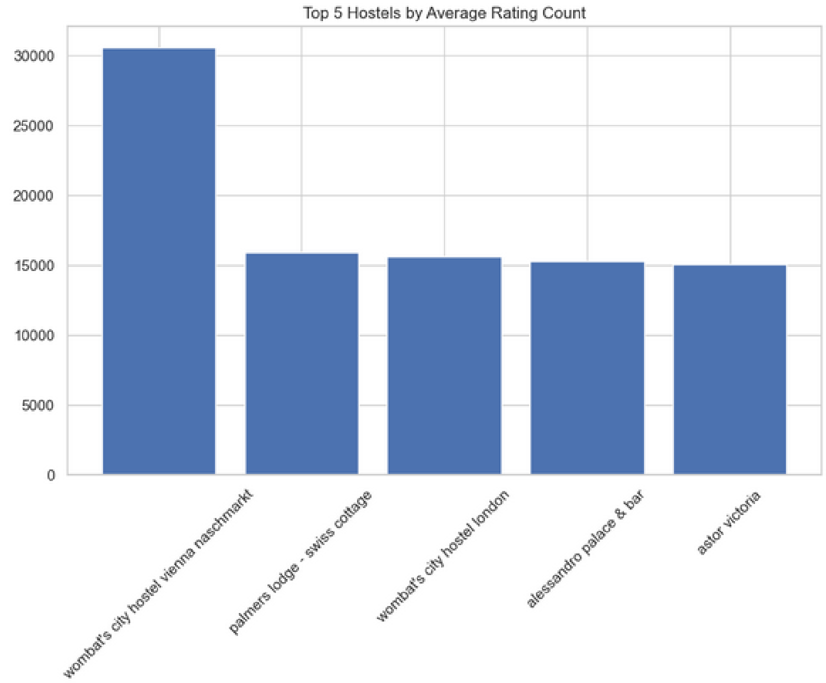
The following is a visual representation of these top cities in barchart dashboard format.



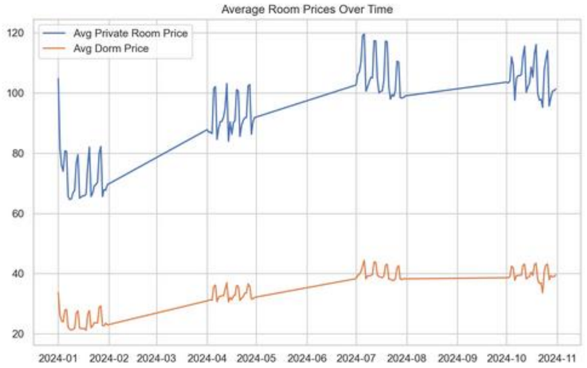
# note: according to the guidelines of the assignment, anything past this point may be additional

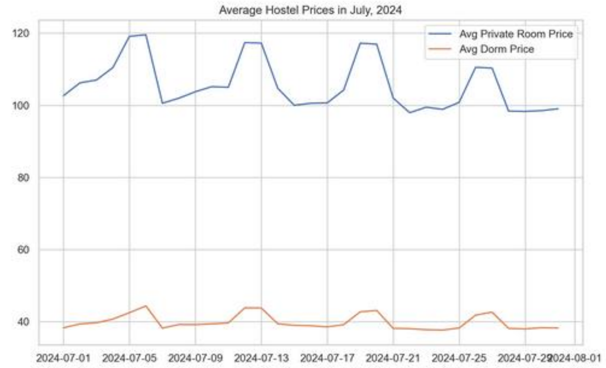
Next, the data was aggregated by hostel name for the purpose of seeing if there are any hostels that tend to have more guests. It was determined that the rating count may represent the popularity of a hostel, or a hostel that tends to have more guests.

- Top 5 Hostels by Average Rating Count
  - Wombat's City Hostel Vienna Naschmarkt (30552)
  - Palmers Lodge - Swiss Cottage (15867)
  - Wombat's City Hostel London (15654)
  - Alessandro Palace & Bar (15273)
  - Astor Victoria (15076)

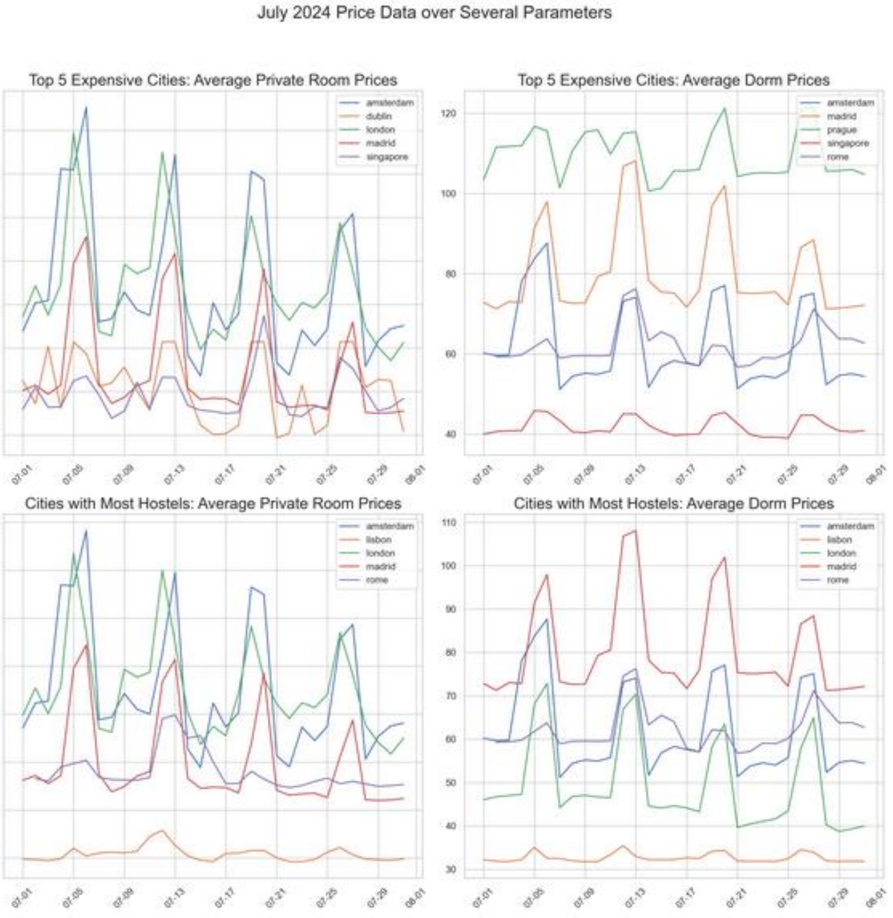


Two plots were created for comparing private room prices and dorm prices across a time series to see how the data is different. In these two graphs, a clear form of seasonality on a small scale is seen by the ups and downs of the data within the month. Additionally, prices seem to get more expensive during the summertime. This was filtered down to the month of July to see what the data looks, and it seems like weekends tend to have an increase in price across the dataset.





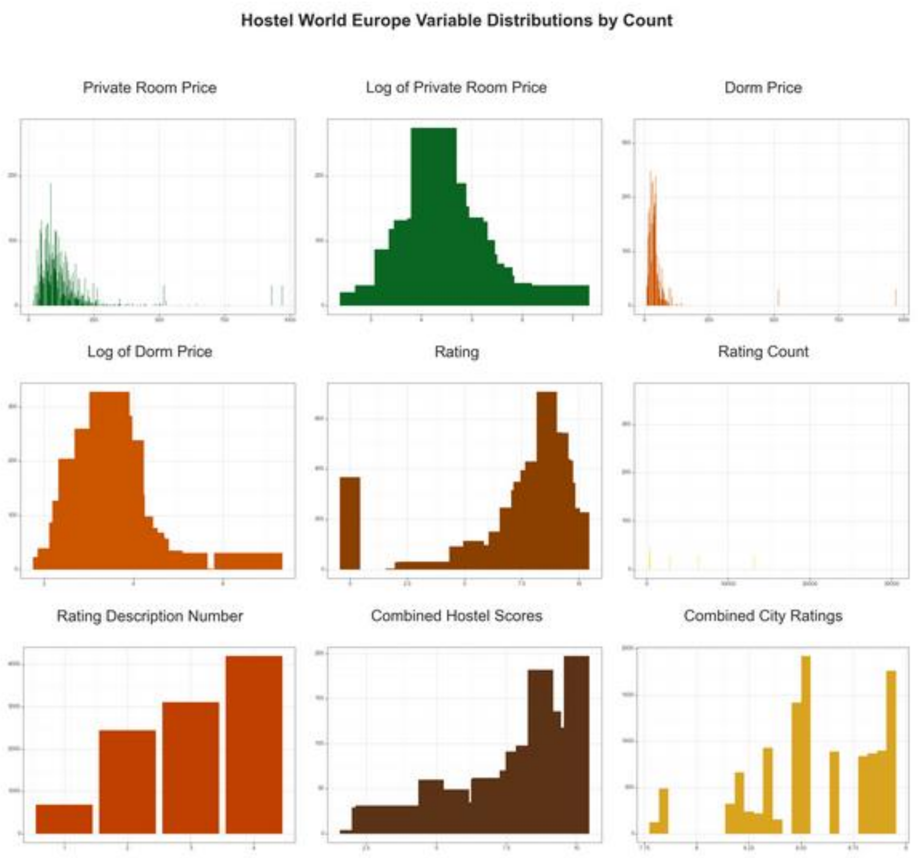
This was further invested by taking a look at the top five most expensive cities for private room prices and dorm prices over July.



A very clear trend in monthly data where there are consistent peaks throughout the month. Most likely this is weekend data. So, this project takes a step further to investigate potential associations between the data collected and prices.

Descriptive Statistics

The data was filtered to July observations in Europe for analysis simplicity. Distribution charts were created to determine how data should be handled. It was determined that prices were in log normal distribution and needed to be log transformed to view the relative price data. Additionally, the detailed city ratings and detailed hostel scores had very repetitive distributions, looking almost identical. They were merged through the means as a result to avoid multicollinearity issues in the regressions. A potential limitation to this is that a proper test to determine if these scores are measuring in the same way was not conducted. The distribution charts can be found here:



Rating count did not have an ideal distribution and was thus dropped from further analysis.

Regression Analysis

This regression analysis will look at what variables have a potential association with the log of price room prices and the log of dorm prices in Europe, July 2024. The cities that are being looked at are the top five cities for most number of hostels by hostel count. This was determined by counting the unique hostel names, and determining which city had the most number of these hostels.

Two analyses will be conducted, one for y = `ln_private_rm_price`, and one for y = `ln_dorm_price`. ln_private_rm_price represents the log transformation of private room prices, while ln_dorm_price represents the log transformation of dorm prices. An OLS model with an interaction model was conducted.

The following will be the explanatory variables for each model to see if any have an association between the prices of dorms or private rooms:

- `Weekend`: binary variable that is 1 if the price was collected from a Friday, Saturday, or Sunday

- `rating`: the overall rating of a hostel

- `rating_descript_num`: an originally qualitative variable that ranked hostels as Good (1), Very Good (2), Fabulous (3), and Superb (4)

- `combined_hostel_scores`: merged by the mean of all individual hostel scores

- `combined_city_ratings`: merged by the mean of all individual city scores

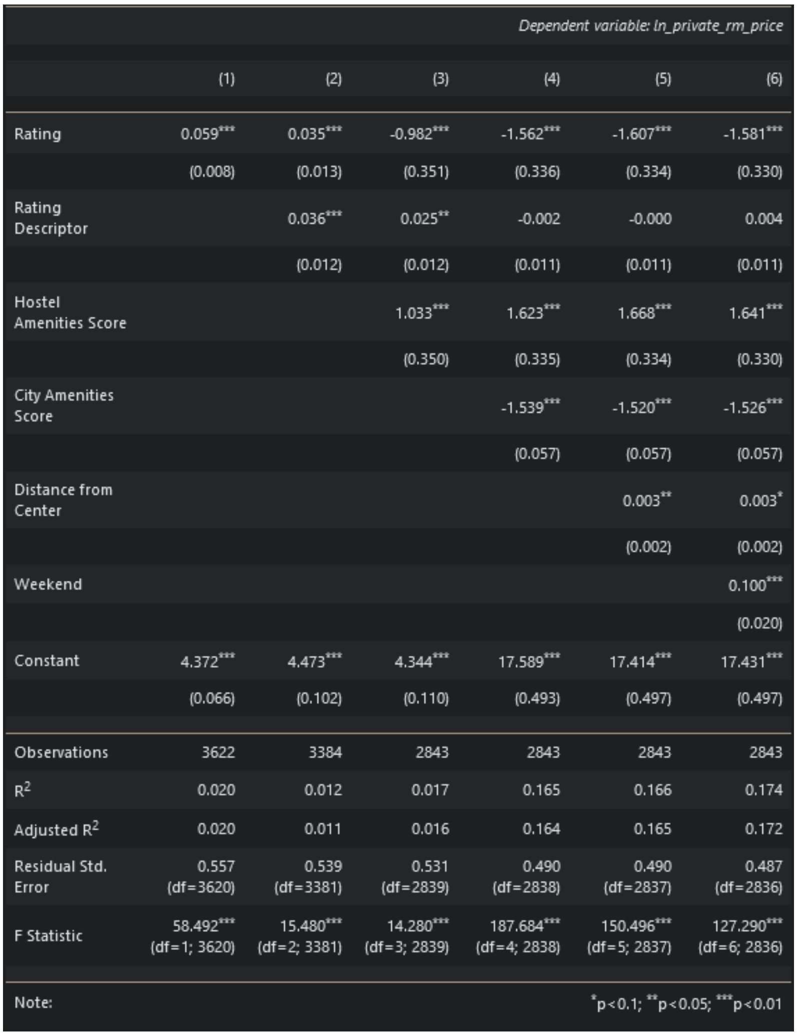
- `distance`: distance from the city center

Post cleaning and after all null values were dropped for the private rooms prices data frame and filtered to top five cities by unique hostel count: n = 3711

Post cleaning and after all null values were dropped for the dorm prices data frame and filtered to top five cities by unique hostel count: n = 3349

Regression analysis on private room prices

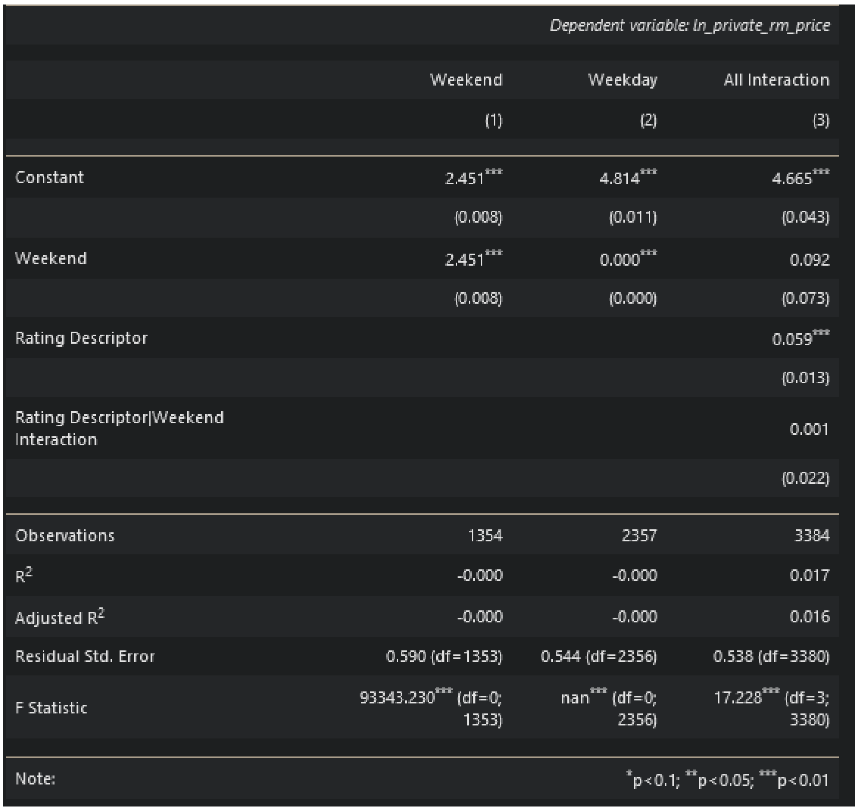
Regression 1: Private room prices with: ratings, rating descriptors, hostel amenities, city amenities, distance from the center, and weekend binary variable.



						Dependent variable: ln_private_rm_price
	(1)	(2)	(3)	(4)	(5)	(6)
Rating	0.059***	0.035***	-0.982***	-1.562***	-1.607***	-1.581***
	(0.008)	(0.013)	(0.351)	(0.336)	(0.334)	(0.330)
Rating Descriptor		0.036***	0.025**	-0.002	-0.000	0.004
		(0.012)	(0.012)	(0.011)	(0.011)	(0.011)
Hostel Amenities Score			1.033***	1.623***	1.668***	1.641***
			(0.350)	(0.335)	(0.334)	(0.330)
City Amenities Score				-1.539***	-1.520***	-1.526***
				(0.057)	(0.057)	(0.057)
Distance from Center					0.003**	0.003*
					(0.002)	(0.002)
Weekend						0.100***
						(0.020)
Constant	4.372***	4.473***	4.344***	17.589***	17.414***	17.431***
	(0.066)	(0.102)	(0.110)	(0.493)	(0.497)	(0.497)
Observations	3622	3384	2843	2843	2843	2843
R^2	0.020	0.012	0.017	0.165	0.166	0.174
Adjusted R^2	0.020	0.011	0.016	0.164	0.165	0.172
Residual Std. Error	0.557 (df=3620)	0.539 (df=3381)	0.531 (df=2839)	0.490 (df=2838)	0.490 (df=2837)	0.487 (df=2836)
F Statistic	58.492*** (df=1; 3620)	15.480*** (df=2; 3381)	14.280*** (df=3; 2839)	187.684*** (df=4; 2838)	150.496*** (df=5; 2837)	127.290*** (df=6; 2836)
Note:						*p<0.1; **p<0.05; ***p<0.01

- Ratings were found to have a statistically significant negative correlation with log prices of private rooms. This could be related to customers will rate a hostel higher when they pay lass for this hostel.
- Hostel Amenities Score was found to have a statistically significant positive correlation with log prices of private rooms. The higher the aggregated hostel amenities scores tend to be associated with higher prices. This may be because the more a hostel puts into the actual hostel itself, the more expensive the cost will be.
- City Amenities Score were found to have a statistically significant negative correlation with log prices of private rooms. This could be related to the fact that individuals enjoy a city more when they paid less for their hostel, since they feel more comfortable spending money on activities in the city.
- Weekends were found to have a statistically significant positive correlation with log prices of private rooms. This could be related to partying on the weekends, or a higher demand in rooms on the weekends.

Regression 2: Private room prices with: weekend, weekday, and interaction



			Dependent variable: ln_private_rm_price
	Weekend	Weekday	All Interaction
	(1)	(2)	(3)
Constant	2.451***	4.814***	4.665***
	(0.008)	(0.011)	(0.043)
Weekend	2.451***	0.000***	0.092
	(0.008)	(0.000)	(0.073)
Rating Descriptor			0.059***
			(0.013)
Rating Descriptor|Weekend Interaction			0.001
			(0.022)
Observations	1354	2357	3384
R^2	-0.000	-0.000	0.017
Adjusted R^2	-0.000	-0.000	0.016
Residual Std. Error	0.590 (df=1353)	0.544 (df=2356)	0.538 (df=3380)
F Statistic	93343.230*** (df=0; 1353)	nan*** (df=0; 2356)	17.228*** (df=3; 3380)
Note:			*p<0.1; **p<0.05; ***p<0.01

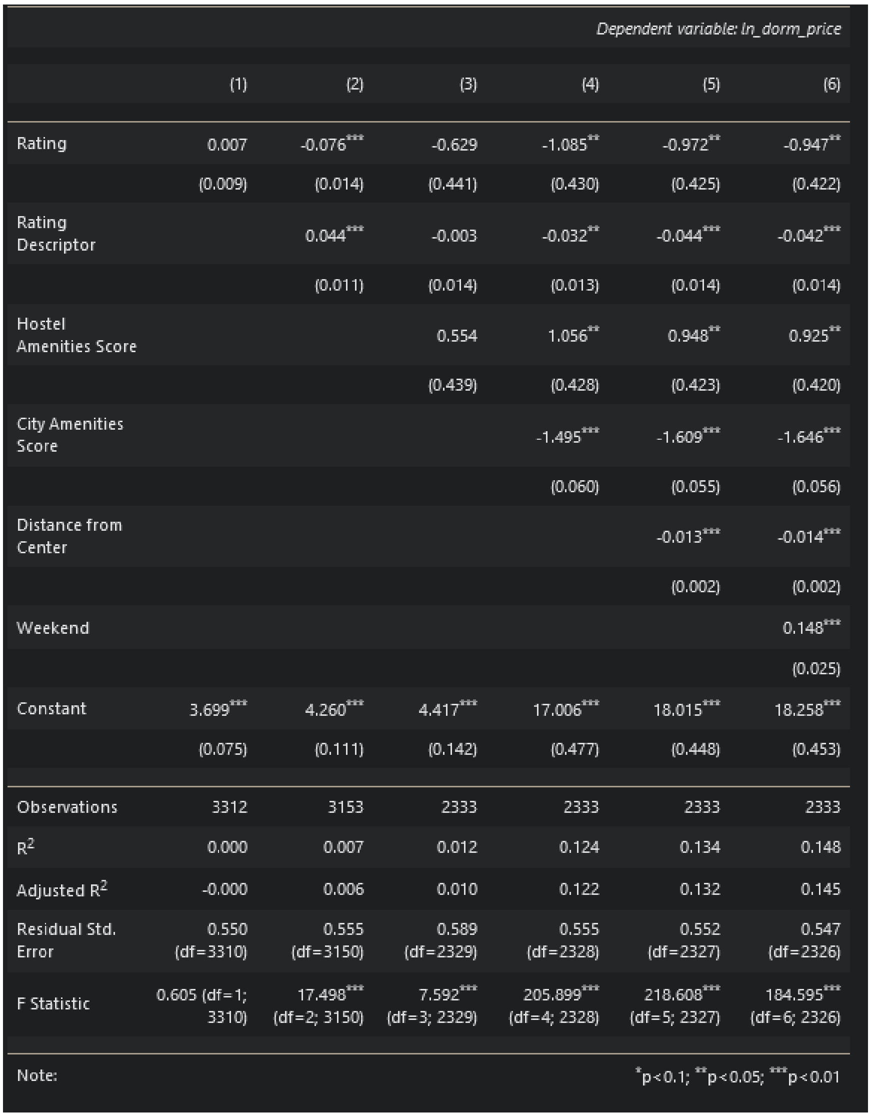
Let's take a closer look at the weekends and weekdays. This regression table looks at them individually, and then incorporates the rating descriptor.

- Unconditionally, weekend private room prices are higher than weekday private room prices. This is statistically significant at the 1% threshold. This suggests that there is a significant in prices over the weekends compared to weekdays.
- Rating Descriptors were found to have a statistically significant positive correlation on weekday private room prices. When interacted with weekends, these descriptors no

longer have any correlation with prices. This would imply that there is a positive association between prices and rating descriptors.

Regression Analysis on Dorm Prices

Regression 3: Dorm prices with: ratings, rating descriptors, hostel amenities, city amenities, distance from the center, and weekend binary variable.



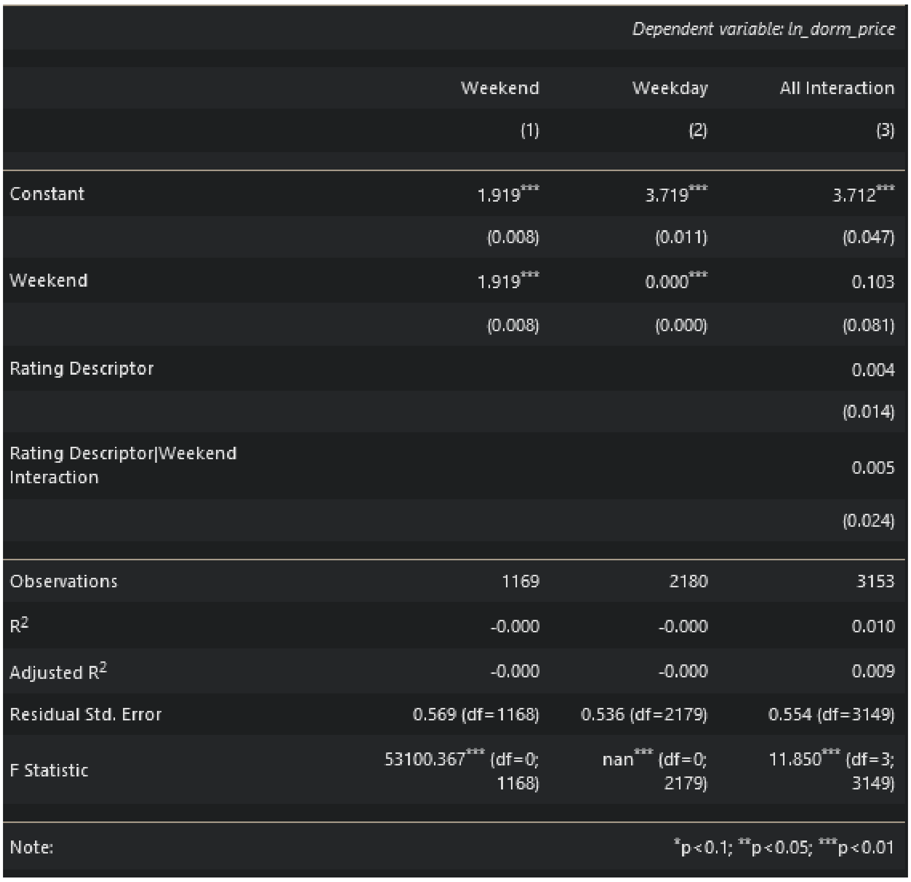
	Dependent variable: ln_dorm_price					
	(1)	(2)	(3)	(4)	(5)	(6)
Rating	0.007	-0.076***	-0.629	-1.085**	-0.972**	-0.947**
	(0.009)	(0.014)	(0.441)	(0.430)	(0.425)	(0.422)
Rating Descriptor		0.044***	-0.003	-0.032**	-0.044***	-0.042***
		(0.011)	(0.014)	(0.013)	(0.014)	(0.014)
Hostel Amenities Score			0.554	1.056**	0.948**	0.925**
			(0.439)	(0.428)	(0.423)	(0.420)
City Amenities Score				-1.495***	-1.609***	-1.646***
				(0.060)	(0.055)	(0.056)
Distance from Center					-0.013***	-0.014***
					(0.002)	(0.002)
Weekend						0.148***
						(0.025)
Constant	3.699***	4.260***	4.417***	17.006***	18.015***	18.258***
	(0.075)	(0.111)	(0.142)	(0.477)	(0.448)	(0.453)
Observations	3312	3153	2333	2333	2333	2333
R^2	0.000	0.007	0.012	0.124	0.134	0.148
Adjusted R^2	-0.000	0.006	0.010	0.122	0.132	0.145
Residual Std. Error	0.550 (df=3310)	0.555 (df=3150)	0.589 (df=2329)	0.555 (df=2328)	0.552 (df=2327)	0.547 (df=2326)
F Statistic	0.605 (df=1; 3310)	17.498*** (df=2; 3150)	7.592*** (df=3; 2329)	205.899*** (df=4; 2328)	218.608*** (df=5; 2327)	184.595*** (df=6; 2326)
Note:						*p<0.1; **p<0.05; ***p<0.01

- Ratings were found to have a statistically significant negative correlation with log prices of dorms. This could be related to customers will rate a hostel higher when they pay lass for this hostel.
- Rating Descriptor (Good, Very Good, Fabulous, Superb) was found to have a statistically significant negative correlation with log prices of private rooms. This could be due to the same reason as the ratings variable.
- Hostel Amenities Score was found to have a statistically significant positive correlation with log prices of dorm room prices. The higher the aggregated hostel

amenities scores tend to be associated with higher prices. This may be since the more a hostel puts into the actual hostel itself, the more expensive the cost will be.

- City Amenities Score were found to have a statistically significant negative correlation with log prices of dorms. This could be related to the fact that individuals enjoy a city more when they paid less for their hostel, since they feel more comfortable spending money on activities in the city.
- Distance was found to have a statistically significant negative correlation with log dorm prices. This could be associated to the fact that the close a hostel is to the centre, the more expensive it will be due to demand.
- Weekends were found to have a statistically significant positive correlation with log prices of dorm rooms. This could be because of partying on the weekends.

Regression 4: Dorm Prices with: weekend, weekday, and interaction



	Dependent variable: ln_dorm_price		
	Weekend	Weekday	All Interaction
	(1)	(2)	(3)
Constant	1.919***	3.719***	3.712***
	(0.008)	(0.011)	(0.047)
Weekend	1.919***	0.000***	0.103
	(0.008)	(0.000)	(0.081)
Rating Descriptor			0.004
			(0.014)
Rating Descriptor|Weekend Interaction			0.005
			(0.024)
Observations	1169	2180	3153
R^2	-0.000	-0.000	0.010
Adjusted R^2	-0.000	-0.000	0.009
Residual Std. Error	0.569 (df=1168)	0.536 (df=2179)	0.554 (df=3149)
F Statistic	53100.367*** (df=0; 1168)	nan*** (df=0; 2179)	11.850*** (df=3; 3149)
Note:			*p<0.1; **p<0.05; ***p<0.01

Unconditionally, weekend private room prices are higher than weekday private room prices. This is statistically significant at the 1% threshold. There is a clear differentiation between prices on the weekends and the weekdays, where weekends are significantly more expensive than weekdays.

Conclusion

When it comes to hostel prices, private rooms and dorm rooms seem to be associated with similar factors. One exception to this was that dorm prices were associated by changes in distance, while private room prices were not. This could be related to the fact that people who seek private rooms may not be as ‘socialable’, and may not want to be near the city centre as much, so prices for private rooms are not impacted by distance, whereas with dorms, there is a chance that hostels have found that people renting dorm spaces are more likely to pay more to be closer to the city centre.

Weekdays were found to consistently have statistically significant higher prices as compared to weekdays. This could be tied to individuals who use Hostel World want to party on the weekends and are willing to pay more money for a weekend hostel.

City Amenities Scores, which include rating scales for how guests feel about a city and the different aspects that city has to offer (such as shopping, eating out, activities, etc.) was consistently found to be significantly negatively correlated to Hostel World prices. This could be since when individuals pay less for their accommodation, they have a greater sentiment for the city because they have more spending capabilities.

Hostels Amenity Scores had a consistently significant positive correlation, suggesting that higher quality hostels are more expensive. However, the overall rating score for the hostel was found to have a significant negative correlation, suggesting that individual will rate a hostel higher when they pay less for this hostel.