
ECON253 DATA ANALYSIS I



ASSESSMENT 2: REPORT
PART 2

2.1 Getting started

```
rm(list=ls())
```

- Removes all the objects from the current workspace to ensure no old data interferes with the new analysis.

```
library(tidyverse)
```

- Loads the `tidyverse` package, which is a collection of R packages designed for data science.

```
library(modelsummary)
```

- Loads the `modelsummary` package for creating tables and summaries of statistical models.

```
library(lsplines)
```

- Loads `lsplines` package, used for linear spline regression modeling.

```
df <- read_csv('hotelbookingdata.csv')
```

- Reads in a CSV file named `hotelbookingdata.csv` and assigns it to the dataframe `df` using `read_csv` from the `readr` package (part of `tidyverse`).

```
df <- rename(df, city = s_city, distance = center1distance)
```

- Renames two columns in `df`:
 - `s_city` becomes `city`
 - `center1distance` becomes `distance`

```
city_data <- filter(df, city == 'Paris', year == 2017 & month == 11 & weekend == 0) | >
```

- Filters the dataframe `df` to select rows where:
 - `city` is "Paris"
 - `year` is 2017
 - `month` is November (`month == 11`) - `weekend` is 0 (likely indicating weekdays).

```
separate(accommodationtype, '@', into = c('garbage', 'acc_type')) | >
```

- Splits the column `accommodationtype` based on the "@" symbol into two new columns:
 - The part before "@" is stored in the column `garbage`.
 - The part after "@" is stored in the column `acc_type`.

```
separate(distance, ' ', into = c('distance', 'miles')) | >
```

- Splits the column `distance` by a space (' ') into two columns:
 - The first part (numeric distance) is kept in `distance`.
 - The second part (likely the word "miles") is stored in `miles`.

```
select(-garbage, -miles) | >
```

- Drops the `garbage` and `miles` columns from the dataframe. These columns were intermediate results from the `separate` operations.

```
filter(acc_type == 'Hotel') | >
```

- Filters the data to keep only rows where the `acc_type` column equals "Hotel", narrowing the data down to hotel accommodations.

```
select(hotel_id, distance, price, neighbourhood, starrating) | >
```

- Selects the following columns to keep:

- `hotel_id`
- `distance`
- `price`
- `neighbourhood` - `starrating`.

`mutate(distance = as.numeric(distance)) | >`





- Converts the `distance` column (which is likely a string or character type) into a numeric data type using `as.numeric()`.

`distinct()`

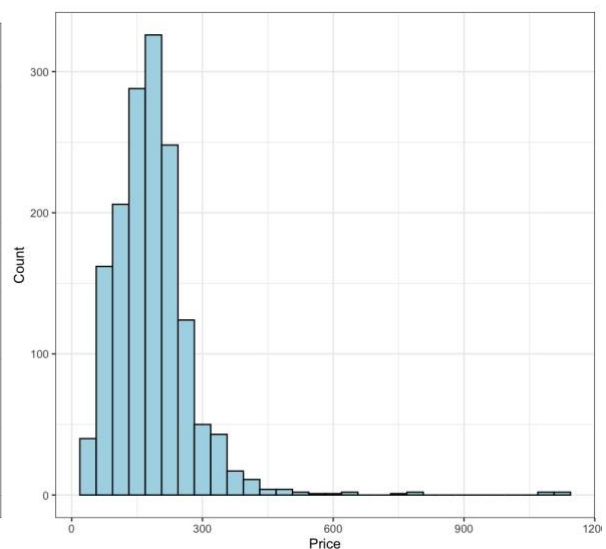
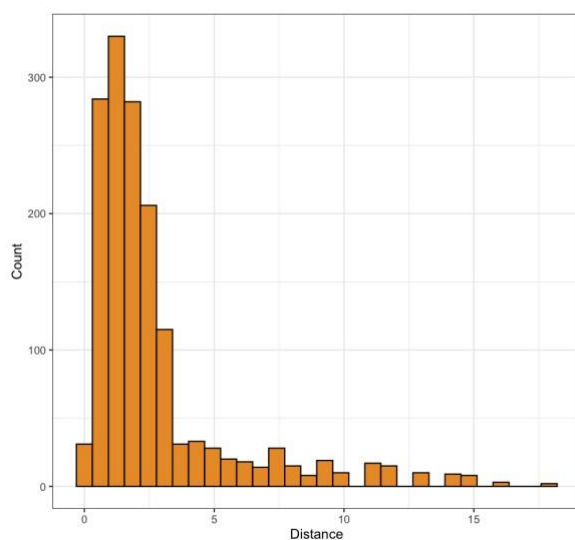
- Removes duplicate rows from the dataframe, keeping only unique rows based on the selected columns.

This script processes hotel booking data, focuses on hotels in Paris in November 2017 (weekdays), cleans up unnecessary columns, and ensures all distance values are numeric for further analysis.

2.2 Explore the Distribution of the Variables

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	Histogram
hotel_id	1536	0	13295.7	644.3	12215.0	13311.5	14398.0	
distance	103	0	2.7	2.9	0.1	1.8	18.0	
price	331	0	184.5	95.6	38.0	177.0	1126.0	
starrating	6	0	2.9	1.3	0.0	3.0	5.0	

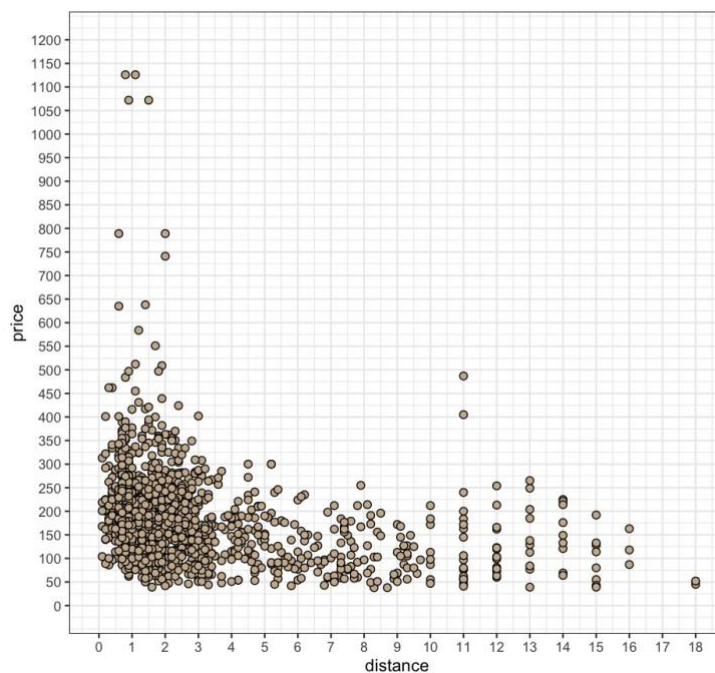
Before starting with specific histograms for price and distance variable, I use a data summary code to have a general look.



The distribution of distance is right-skewed (positively skewed). Most of the observations are concentrated near 0, with a sharp drop-off as the distance increases. There are fewer observations at higher distances, showing that most hotels in the dataset are located close to the central point (likely the city centre), and when distance increases, the number of hotels significantly decreases.

Similarly, the distribution of price is also right-skewed (positively skewed). A large number of hotels are priced in the lower range (around 0-300), while very few have higher prices, shown by the tapering tail toward the right of the distribution. This indicates that most hotels have relatively lower prices, and there are only a few higher-priced hotels.

(ii)



For distance, there are a few hotels located at relatively far distances from the centre (around 10–18). These could be considered outliers when compared to the majority of hotels concentrated closer to the centre. These outlying hotels might need further investigation to understand why they are located far from the rest.

For price, there are extreme values in the distribution as well. Some hotels are priced as high as 1200, which is much higher than the majority (below 300). These extreme values are clearly outliers and may represent luxury or premium hotels that are significantly more expensive than the average.

(iii)

The scatterplot of distance vs price reveals a non-linear and inverse relationship. As distance increases, the price tends to decrease. In particular:

- Most hotels located near the centre with a wide range of price (both high and low price)

- As distance increases, price tends to stabilize at a lower price, with a few high-priced hotels
- There is a concentration of low price hotels (50-450), especially at distance from 0-5
- There seems to be a dense cluster of hotels that are both close to the centre and lowerpriced (around 0-5 km and prices between 50-300).

2.3 Addressing Extreme Values

The purpose of the analysis, as assumed, is to focus on regular customers, who seeks for regular accommodations with affordable price (150 – 300) with reasonable reviews (hotels ranged from 3-4 stars), relatively close to facilities and city centre (less than 10km from city centre). Therefore, I develop a set of criteria based on which extreme values can be handled effectively:

1. Price Criteria

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	Histogram
hotel_id	1536	0	13295.7	644.3	12215.0	13311.5	14398.0	
distance	103	0	2.7	2.9	0.1	1.8	18.0	
price	331	0	184.5	95.6	38.0	177.0	1126.0	
starrating	6	0	2.9	1.3	0.0	3.0	5.0	

Statistically, we can look at the standard deviation (SD) to decide at which threshold we can drop those observations. In this case, the price variable has the SD of 95.6, which means on average, hotel prices vary by about 95.6 from the mean price of 184.5. As a rule of thumb, outliers are ± 3 times the SD, which can be interpreted as 2 to 3 times the standard deviation away from the mean is considered an outlier. In this case, prices above 400 are considered outliers because they are more than 2 times the SD above the mean price. Therefore, choosing 400 dollars as a threshold to drop the outliers is an optimal option.

Hotels with prices above 400 are often luxury hotels with high star ratings or other unique characteristics that make them outliers in the dataset. But in this case, star rating is a factor that has strong influence on price.

hotel_id	distance	price	neighbourhood	starrating	15	13362	0.4	462	Louvre – Place Vendome (1 arr. and 2 arr.)	5
1	12580	1.1	1126 Champs Elysees (8 arr.)	0	16	13391	0.2	401	Louvre – Place Vendome (1 arr. and 2 arr.)	5
2	12593	1.5	1072 Champs Elysees (8 arr.)	5	17	13397	0.6	401	Louvre – Place Vendome (1 arr. and 2 arr.)	5
3	12609	1.1	455 Champs Elysees (8 arr.)	5	18	13419	0.3	462	Louvre – Place Vendome (1 arr. and 2 arr.)	5
4	12625	1.4	638 Champs Elysees (8 arr.)	5	19	13438	0.9	497	Louvre – Place Vendome (1 arr. and 2 arr.)	5
5	12642	1.4	417 Champs Elysees (8 arr.)	4	20	13483	1.0	416	Marais – Pompidou – Notre Dame de Paris (3 arr. and ...	4
6	12655	2.0	789 Champs Elysees (8 arr.)	5	21	13504	1.1	512	Marais – Pompidou – Notre Dame de Paris (3 arr. and ...	4
7	12670	1.2	431 Champs Elysees (8 arr.)	5	22	13600	1.8	497	Montmartre – Sacre Coeur (18 arr.)	0
8	12680	2.0	741 Champs Elysees (8 arr.)	5	23	13681	1.7	551	Montparnasse Tower (14 arr.)	2
9	12681	1.9	439 Champs Elysees (8 arr.)	5	24	13997	11.0	487	Palace of Versailles	4
10	12710	1.9	509 Champs Elysees (8 arr.)	5	25	14096	1.5	421	Paris	4
11	12717	1.2	584 Champs Elysees (8 arr.)	5	26	14223	0.6	789	Saint-Germain-des-Pres	5
12	12859	0.8	484 Eiffel Tower – Orsay Museum (7 arr.)	5	27	14243	0.6	635	Saint-Germain-des-Pres	5
13	13252	0.9	1072 Latin Quarter – Pantheon (5 arr.)	4	28	14305	2.4	424	Trocadero (16 arr.)	5
14	13360	0.8	1126 Louvre – Place Vendome (1 arr. and 2 arr.)	5	29	14326	3.0	402	Trocadero (16 arr.)	0
15	13362	0.4	462 Louvre – Place Vendome (1 arr. and 2 arr.)	5	30	14370	11.0	405	Villepin	4

When I filter out all observations with price higher than 400, I found out that all hotels that have extravagant price are not only because they are close to city centre but are also associated with high star rating (5 star). While keeping these observations can help the analysis capture the luxury niche of hotels, I want this analysis to pay more focus on the average customers, who are looking for average hotel price.

As these extremely high prices do not reflect the typical pricing structure of hotels in Paris, removing these extreme values helps focus the analysis on the more representative hotel market. All hotels with a price exceeding 400 will be dropped from the dataset. 0 star (no rating) and 5 stars (luxury) will essentially be dropped because they don't represent the general customer demand. Similarly, all hotels with star rating lower than 3 will be dropped because the analysis aims to examine the general relationship between distance and price, so keeping other factors such as stars rating constant help address the relationship between the two variables of interest clearer. This action helps providing insights into the broader market, statistically and practically rather than skewing the results with high-end luxury pricing.

2. Distance Criteria

To able to choose the threshold, I examined the patterns of hotels with far distance:

hotel_id	distance	price	neighbourhood	starrating
1	12741	11	Chilly-Mazarin	0
2	12742	11	Chilly-Mazarin	2
3	12895	18	Elancourt	1
4	12896	13	Epinay-sur-Orge	2
5	12897	13	Epinay-sur-Orge	3
6	12898	15	Fleury-Merogis	0
7	12899	15	Fleury-Merogis	1
8	13145	12	Herblay	2
9	13146	12	Herblay	1
10	13312	16	Le Mesnil-Amelot	3
11	13313	15	Le Mesnil-Amelot	4
12	13314	15	Le Mesnil-Amelot	3
13	13315	15	Le Mesnil-Amelot	3
14	13316	15	Le Mesnil-Amelot	0
15	13320	11	Le Thillay	0
16	13337	11	Livry-Gargan	2

17	13468	15	133	Maffliers	4
18	13563	16	118	Mauregard	4
19	13564	16	163	Mauregard	4
20	13565	15	39	Mery-sur-Oise	0
21	13570	13	113	Mitry-Mory	3
22	13762	11	80	Morangis	3
23	13983	11	240	Palace of Versailles	4
24	13986	11	184	Palace of Versailles	3
25	13987	11	145	Palace of Versailles	3
26	13990	11	200	Palace of Versailles	4
27	13992	11	106	Palace of Versailles	2
28	13994	11	163	Palace of Versailles	3
29	13997	11	487	Palace of Versailles	4
30	14149	13	204	Roissy-en-France	4
31	14152	12	162	Roissy-en-France	4
32	14154	13	129	Roissy-en-France	3





33	14155	13	138	Roissy-en-France	4
34	14157	13	185	Roissy-en-France	4
35	14159	12	81	Roissy-en-France	2
36	14161	12	63	Roissy-en-France	1
37	14162	12	120	Roissy-en-France	3
38	14164	12	254	Roissy-en-France	4
39	14165	13	249	Roissy-en-France	4
40	14169	14	121	Roissy-en-France	3
41	14171	14	225	Roissy-en-France	4
42	14172	12	78	Roissy-en-France	2
43	14173	12	103	Roissy-en-France	2
44	14174	12	123	Roissy-en-France	3
45	14175	13	265	Roissy-en-France	4
46	14197	12	213	Saclay	4
47	14203	14	133	Saint-Cyr-l'École	3
48	14204	14	69	Saint-Cyr-l'École	0

49	14208	13	84	Saint-Cyr-l'École	3
50	14209	14	64	Saint-Cyr-l'École	1
51	14284	18	52	Trappes	2
52	14286	14	176	Tremblay-en-France	4
53	14287	14	221	Tremblay-en-France	4
54	14288	14	149	Tremblay-en-France	3
55	14289	14	214	Tremblay-en-France	4
56	14362	11	46	Vigneux-sur-Seine	1
57	14363	11	56	Vigneux-sur-Seine	2
58	14367	12	166	Villepinte	3
59	14368	12	99	Villepinte	1
60	14369	12	103	Villepinte	0
61	14370	11	405	Villepinte	4
62	14371	11	172	Villepinte	2
63	14373	12	122	Villepinte	2
64	14374	11	41	Villepinte	1

Interestingly, I found that hotels located far from the city centre (beyond 10 km) tend to have irregular pricing patterns. With some further research with Google Map, I also found out that most of those far-distance hotels with high price are famous places or close to the airports, for instance:

- Mauregard
- Epinay-sur-Orge
- Le Mesnil-Amelot
- Le Thillay
- Mitry-Mory
- Morangis
- Roissy-en-France
- Tremblay-en-France
- Villepinte

Because the analysis focuses on general customer who is looking for reasonable distance to the city centre, for example, when choosing hotels to stay for trips, customers tend to choose the ones that are not too far from the city centre because they want to save time travelling to tourists spots; and for the purpose of examining the relationship between price and distance, I will not let geological reasons to be a factor here. It is reasonable to identify those hotels which are located more than 10 km away from the city are outliers and need to be dropped.

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	Histogram
hotel_id	1536	0	13295.7	644.3	12215.0	13311.5	14398.0	
distance	103	0	2.7	2.9	0.1	1.8	18.0	
price	331	0	184.5	95.6	38.0	177.0	1126.0	
starrating	6	0	2.9	1.3	0.0	3.0	5.0	





Statistically, to decide the threshold where can we drop the observations that are considered outliers, we can again look at the SD. As for distance, the SD is 2.9, which means the distances from the city centre tend to vary by about 2.9 km from the mean distance of 2.7 km. Based on the rule of thumb mentioned above, distance beyond the range of 9 km ($2.9 \times 3 = 8.7$). However, to better align with the data and geological reasons, I will set a threshold of 10 to drop the outliers.

Based on geological and statistical rationale, action for this will be excluding those observations above 10km from the analysis. This restriction allows the focus to remain on the general city hotel market, where proximity to central tourist attractions and the city centre has a stronger influence on pricing, without interference from hotels that benefit from being near airports or far-off landmarks.

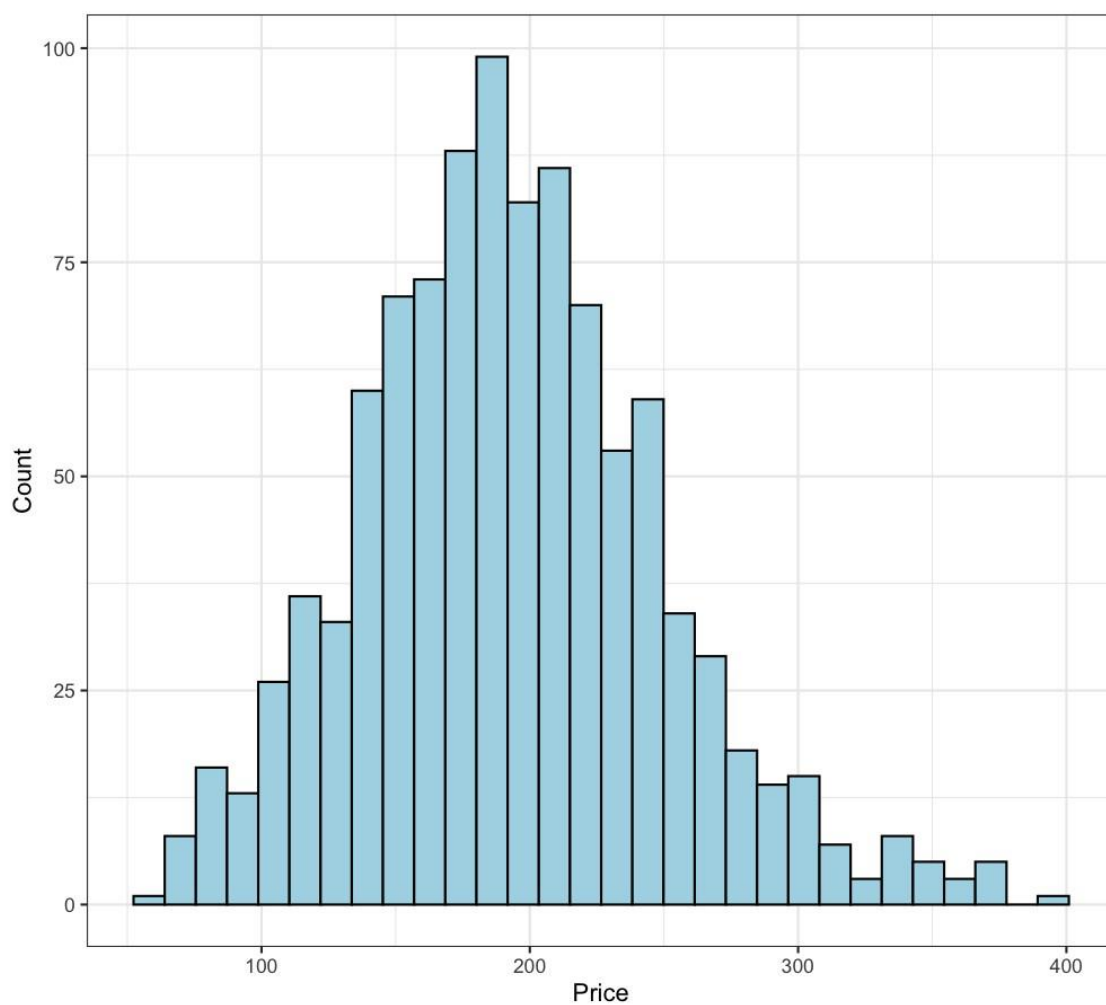
By implementing these criteria, the analysis will better represent the general relationship between hotel prices and distance from the city centre, while avoiding distortion from extreme or atypical cases. This approach ensures that the model provides a clear view of market trends, with a focus on the majority of hotels that cater to typical travellers rather than luxury or niche markets. After finalizing the dataset, histograms and scatterplots will be

rechecked to confirm that the extreme values have been effectively removed and that the remaining data accurately reflect the general trends.

This is an overall look of the data after handling with extreme values:

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	Histogram
hotel_id	1016	0	13271.1	639.6	12215.0	13279.0	14397.0	
distance	89	0	2.2	1.8	0.1	1.7	9.7	
price	238	0	193.4	56.1	53.0	190.0	390.0	
starrating	2	0	3.4	0.5	3.0	3.0	4.0	

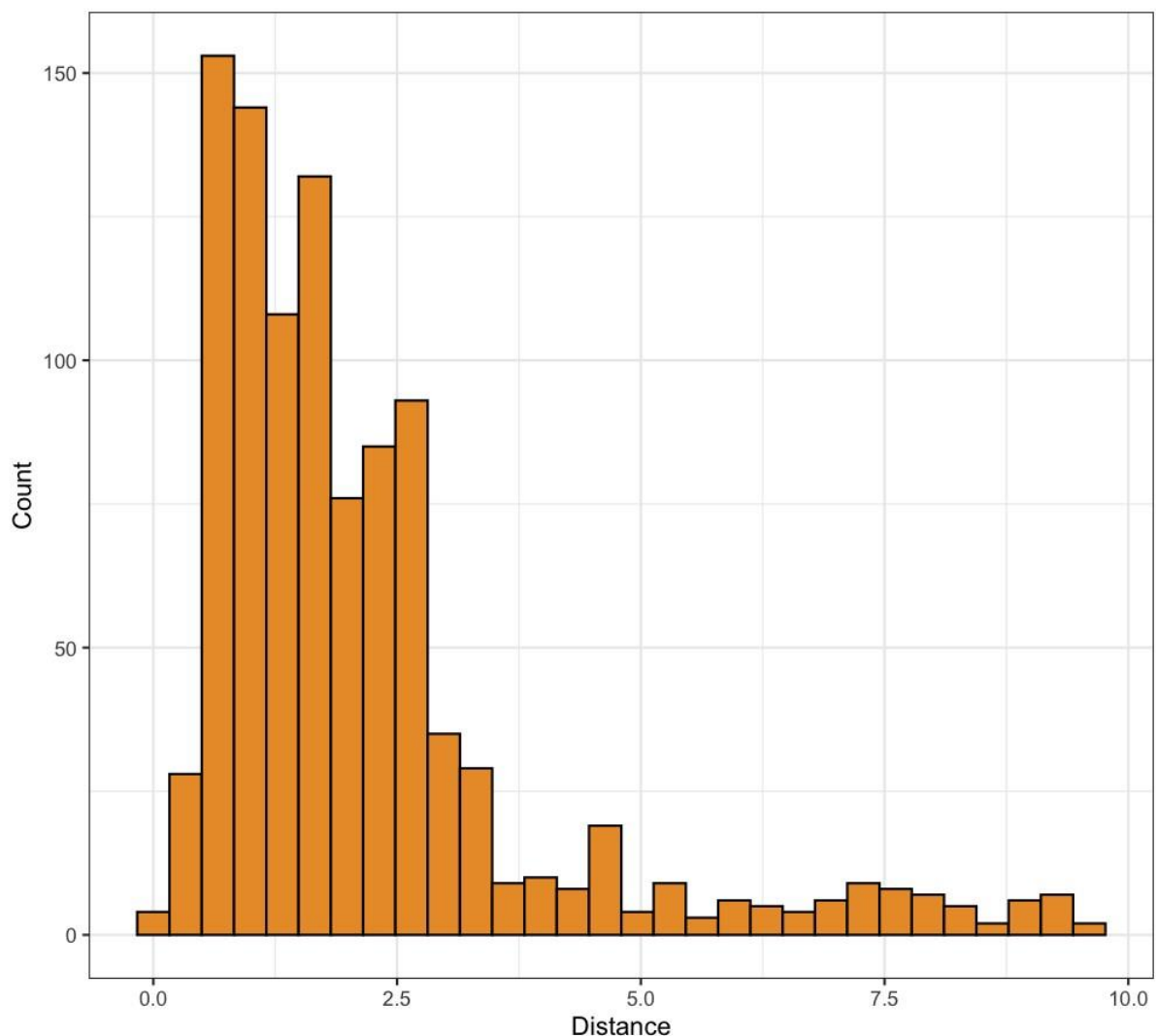
Histogram of price after dropping outliers:



The mean and the median of price variable, respectively are 193.4 and 190, are relatively close, which means that the distributions of price are fairly symmetrical after dropping extreme values.

The standard deviation of 56.1 suggests that there is still some variability in hotel prices, but it is within a more reasonable range compared to the previous (95.6). This indicates that the price spread has been reduced compared to the original dataset, meaning that most of the remaining hotels are priced similarly without extreme outliers.

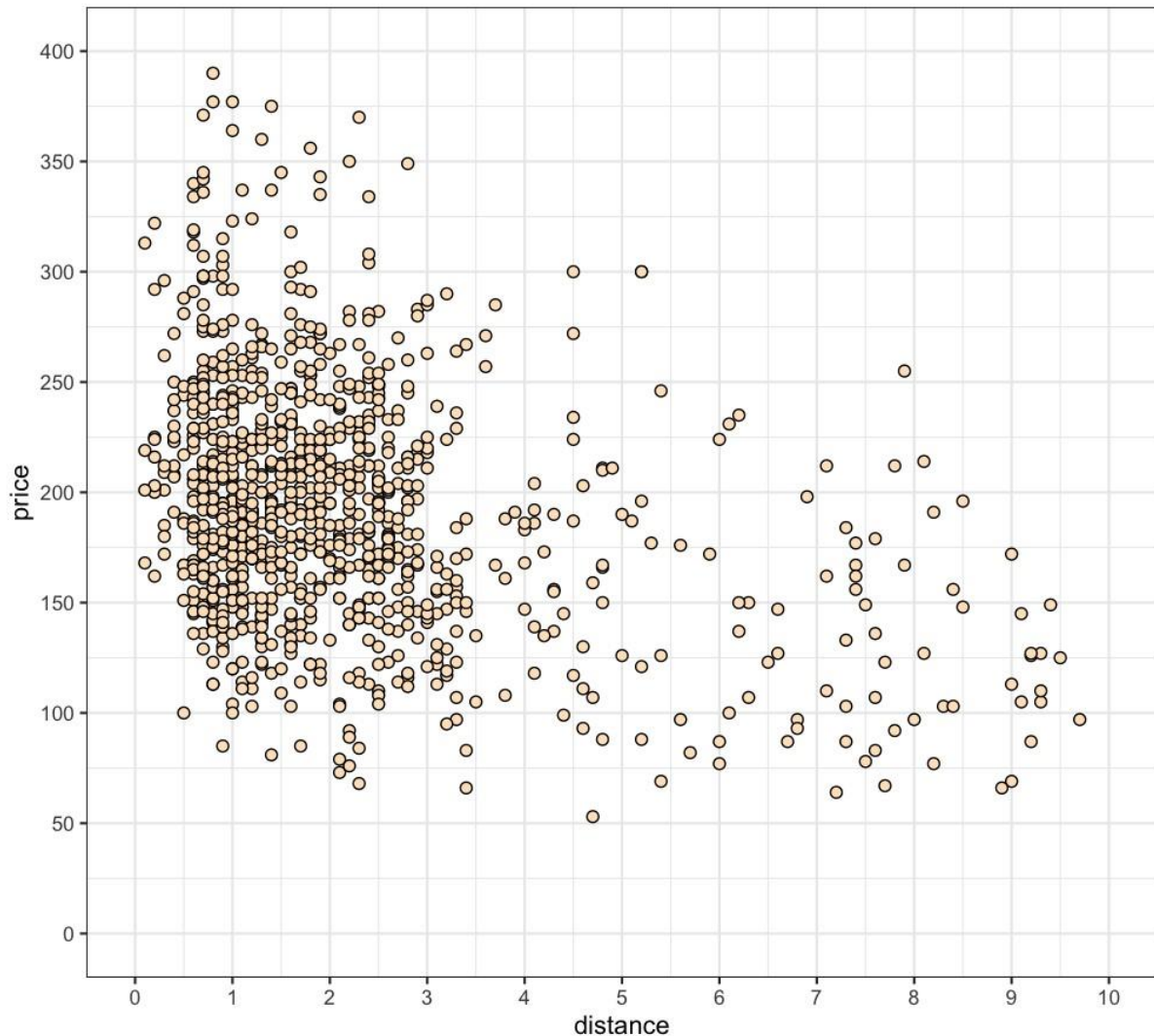
Histogram of distance after dropping outliers:



The mean distance of 2.2 km and median distance of 1.7 km suggest that most hotels are located fairly close to the city centre. The median being slightly lower than the mean could indicate a slight right-skew, meaning a few hotels are located farther away from the city centre, but not to the extreme extent observed in the original dataset.

The standard deviation of 1.8 km indicates moderate variation in hotel locations compared to the previous (2.9 km), with the bulk of hotels clustered within a few kilometres of the centre.

Below is the scatter plot after remove extreme values:



The scatterplot reflects a cleaner distribution, with fewer extreme outliers, especially on the price axis. In the previous version of the data, high-priced hotels likely caused more scatter and noise, distorting the general trend. Now, after eliminating these extremes, the relationship between price and distance is more evident and easier to model.

2.4 Linear regression

a.



The trend depicted by the Lowess curve is non-linear. Even though the overall trend of the curve is downward, there is some fluctuations with rises and falls rather than a straight line

Near the city centre (0 to 2 km), the curve initially shows a sharp decline in hotel prices. This suggests that proximity to the city centre significantly influences hotel pricing, with prices being notably higher for hotels located closer to the city centre. This is likely due to the premium placed on proximity to tourist attractions, business centres, and high-demand areas. However, beyond 2 km, the price decline becomes more gradual, indicating that distance has a diminishing impact on price as you move farther away from the city centre. This indicates that hotels farther out are still affected by distance, but other factors may also play a role. After about 3 km, the slope of the curve flattens slightly, suggesting that prices stabilize as hotels move into less central areas. From 6 to 10 km, there is a relatively flat segment of the curve, suggesting that at greater distances (6+ km), prices don't drop as drastically, reflecting a plateau where hotels further out are priced similarly.

As the curve is non-linear, particularly the sharp drop in prices near the city centre and the plateau beyond 6 km; a simple linear model would not be appropriate for this dataset. If we apply a linear model in this case, it would force a constant rate of price decreases as distance increases, which clearly does not align with the observed data.

(b) *Fitting a Linear Regression Model*

Coefficient	Estimate	Std. Error	T value
Intercept	216.740	2.578	84.09
Slope	-10.739	0.912	-11.78

Multiple R squared	0.12
Adjusted R squared	0.119

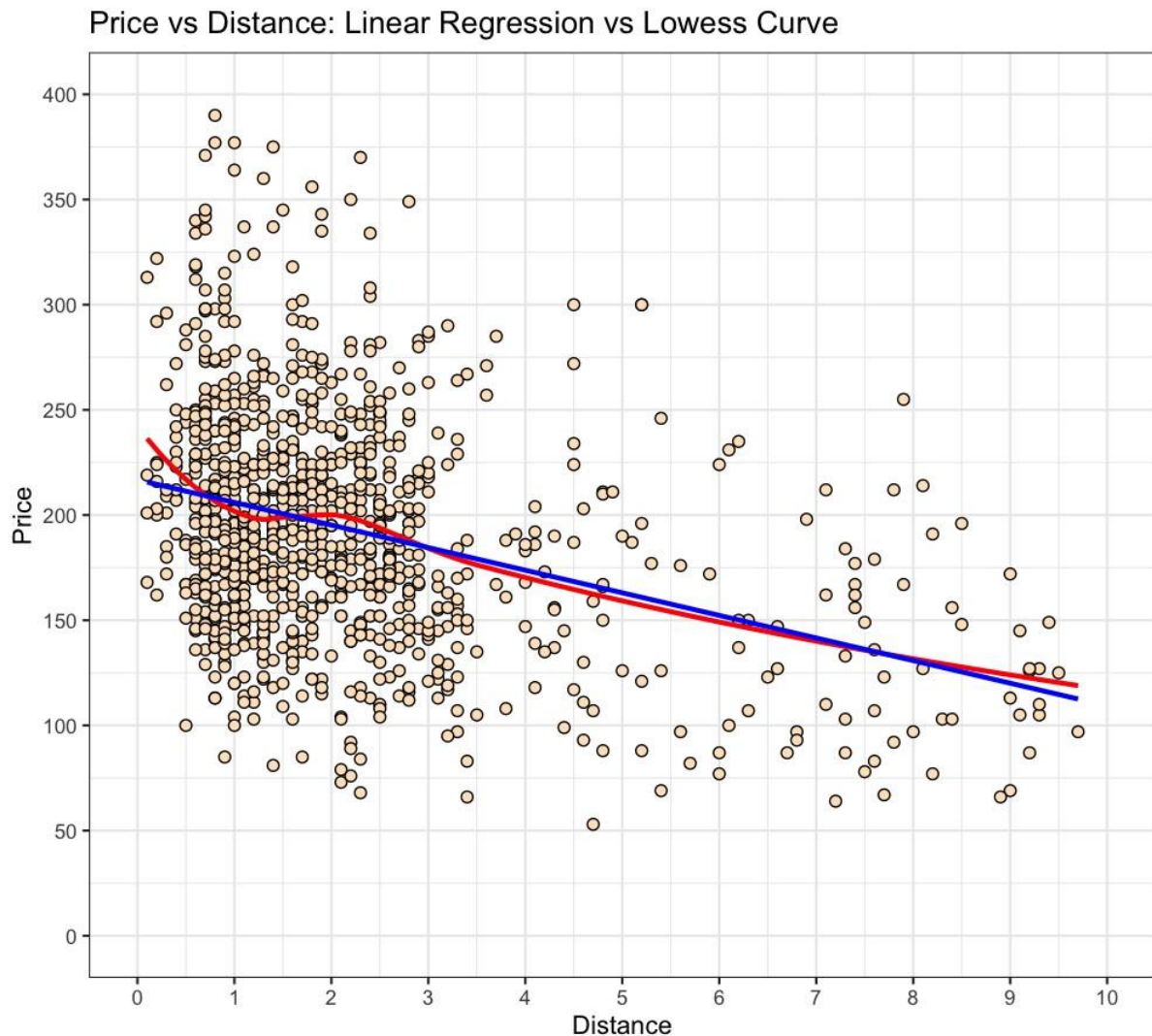
Intercept (216.740):

- The intercept represents the estimated price when the distance from the city centre is zero (for example when a hotel located directly in the city centre).
- With a value of 216.74, this means that the average price for hotels located at a 0 km distance from the city centre is approximately 216.74 units.
- The t-value for the intercept is 84.09, which is very large. This indicates that the estimate of the intercept is 84.09 standard deviations away from zero, meaning the average price of hotels located at the city centre (distance = 0) is significantly different from zero.
- A t-value this large indicates that the intercept is highly reliable and has a strong influence on the model. The fact that it's so far from zero confirms that the estimated starting price is statistically reliable.

Slope for Distance (-10.739):

- The coefficient for distance is -10.739, meaning that for every 1 km increase in distance from the city centre, the hotel price decreases by approximately 10.739 units.
- This negative relationship confirms the intuitive expectation that hotels farther from the city centre tend to have lower prices.
- The t-value for the distance coefficient is -11.78, which is also very large. This means that the coefficient estimate of -10.739 (the effect of distance on price) is 11.78 standard deviations away from zero. The negative sign indicates that as distance increases, price decreases.
- A t-value with a large magnitude (either positive or negative) like this suggests that distance is a strong predictor of hotel price, and this relationship is not likely because of chance.
- Typically, a t-value above 2 or below -2 is considered statistically significant, but here, the magnitude of 11.78 shows an even stronger evidence that distance plays an important role in determining hotel prices.

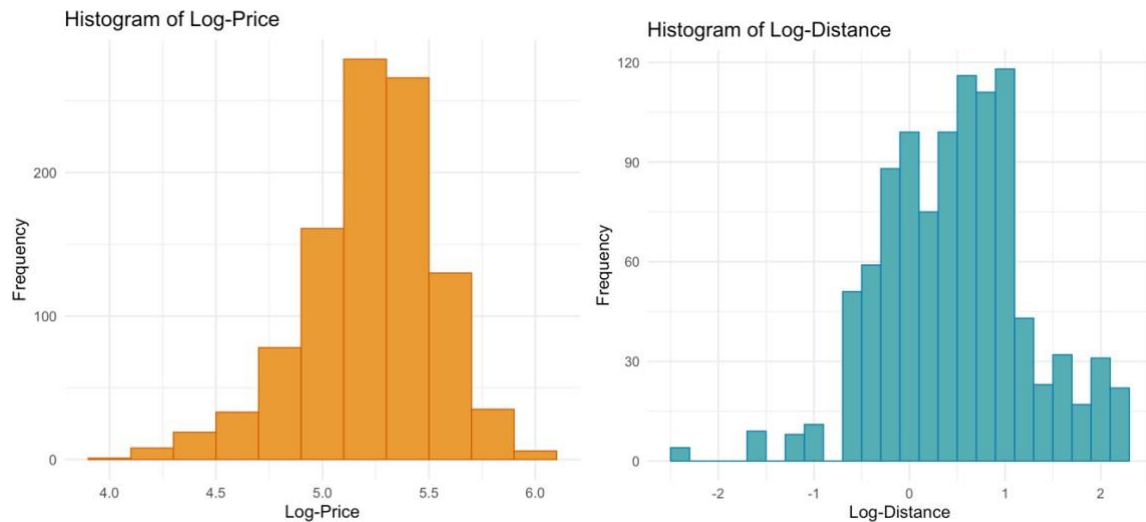
R-squared (0.1203): This indicates that the model explains approximately 12.03% of the variance in hotel prices based on distance. While this shows that distance has some predictive power, a large portion of the price variability is explained by other factors not captured in this model such as the proximity of hotels to airport or major landmarks and star rating, two of which we decide to keep out to only examine the relationship between price and distance.



The plot highlights the limitations of a linear model in capturing the true relationship between hotel price and distance. The Lowess curve provides a more accurate representation, showing that prices drop rapidly near the city centre and then stabilize as distance increases. In this case, the linear model overestimates the effect of distance at greater ranges and underestimates it near the city centre.

2.5 Log transformations

(a) Creating Log-Transformed Variables

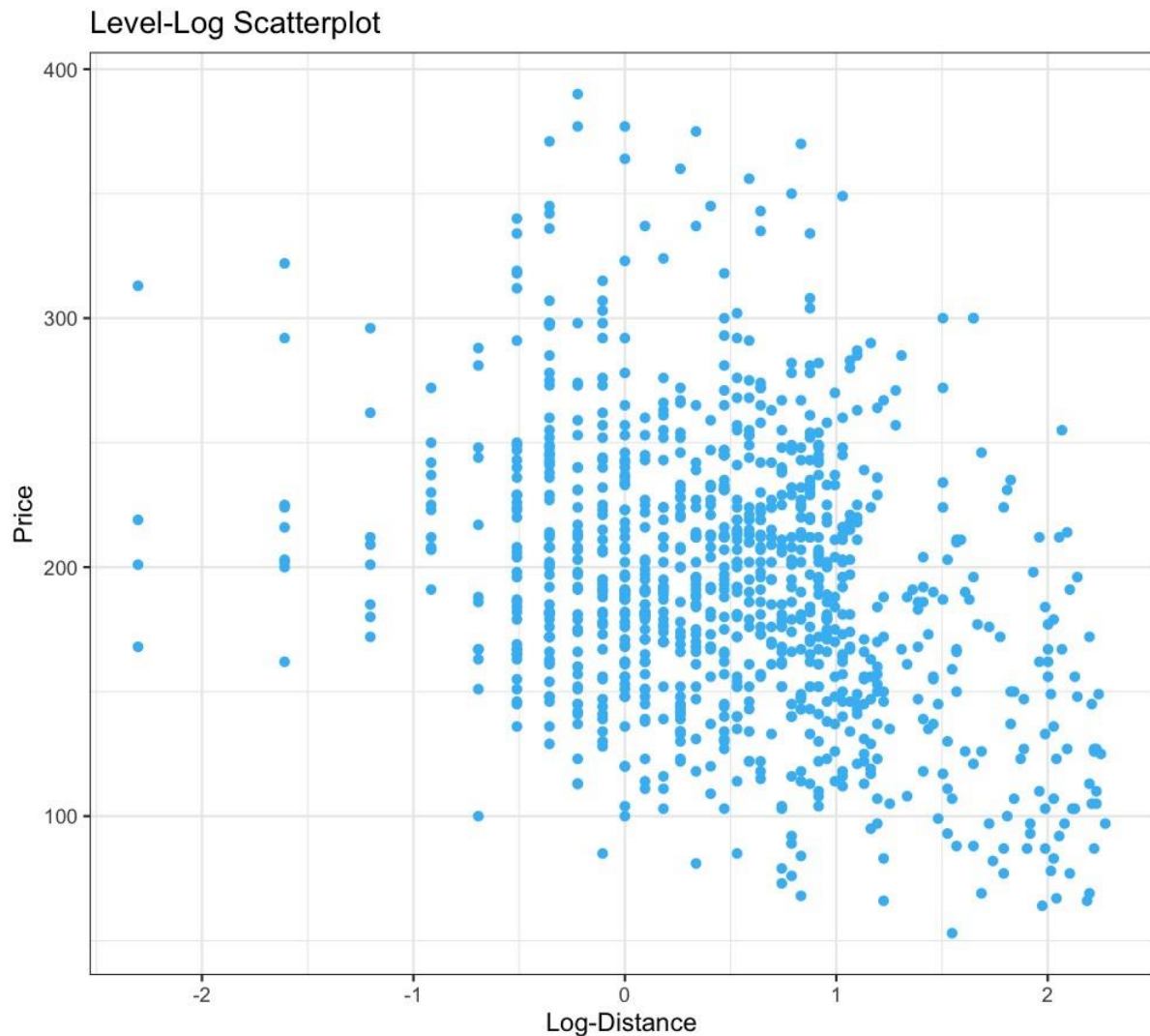


The histogram for log-price shows a relatively symmetric, bell-shaped distribution centred around values between 5 and 5.5. Compared to the original price distribution from section 2.2, which was likely right-skewed, with some extreme high prices, the log transformation has helped to make the data more normally distributed, pulling in the tail of high prices. Also, the mean and median of the log-price are now closer (5.22 and 5.24 respectively), indicating that the log transformation reduced the impact of extreme outliers and skewness in the data. In conclusion, the transformation has been useful in making the price data more evenly distributed, bringing it closer to a normal distribution, which is helpful for statistical analysis.

The histogram for log-distance is somewhat right-skewed, with most values clustering around 0 to 1. However, there are some observations with log-distance values below 0, corresponding to very short distances, and a few observations greater than 2. Compared to the original distance distribution, this log-transformation appears to have helped reduce the rightskewness to some extent, making the bulk of the data more concentrated (mean and median are 0.49 and 0.53 respectively) and less spread out. However, there are still some extreme values in the lower range of the log-distance scale, corresponding to hotels very close to the city centre (which were originally represented by small distances before transformation). These points might still be considered outliers but are now less extreme. In hindsight, while the log transformation has helped to reduce some of the skewness in the distance data, it has not fully normalized the distribution. However, it has made the data less skewed compared to the original form.

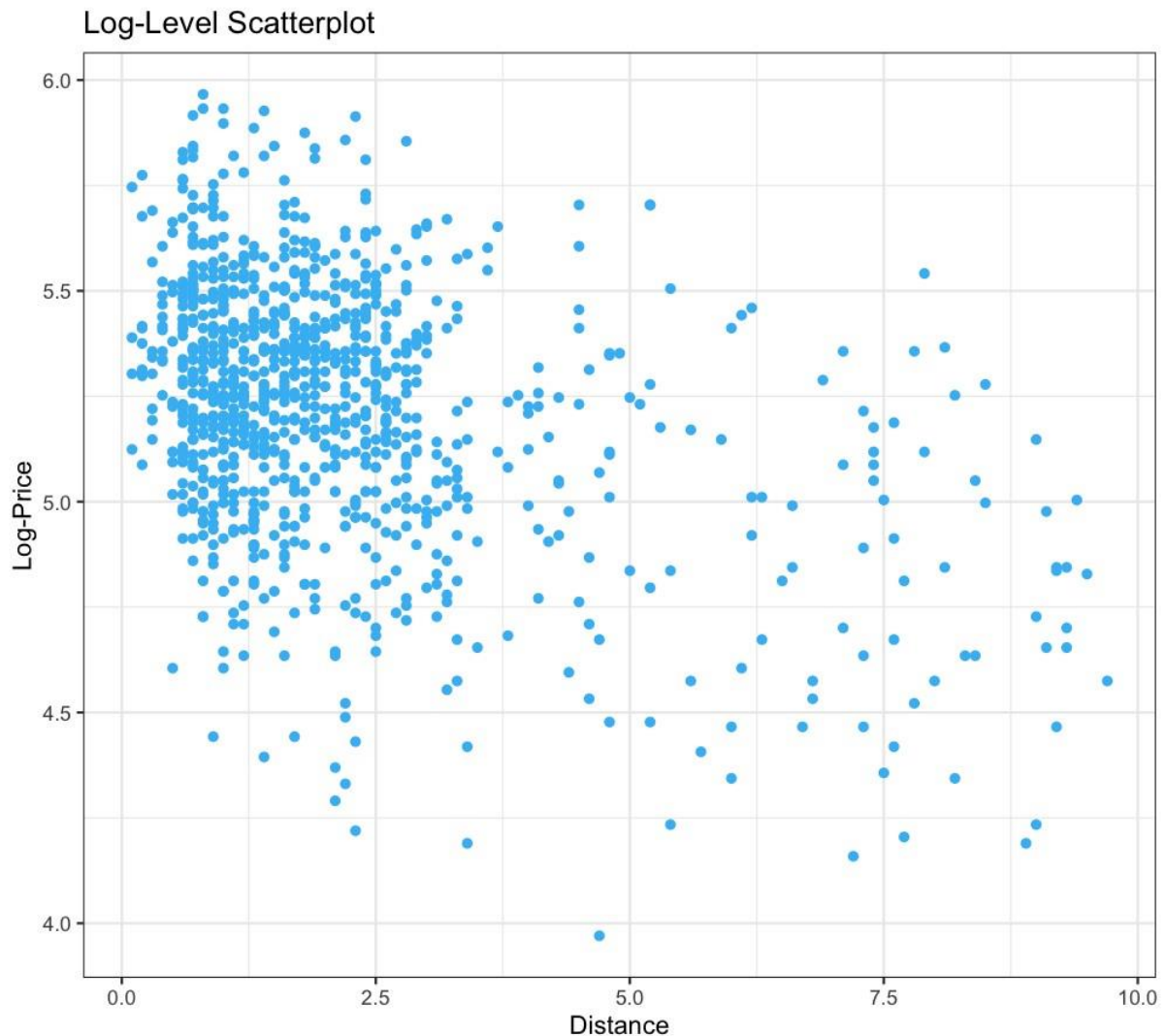
In conclusion, the log transformations have clearly been useful in both cases in making the variables more evenly distributed. By compressing the high values and spreading out the smaller values, bringing both distributions closer to normality, this helps improving the high level of right-skewness of the original histograms for both price and distance. (b)

Scatterplots and model choice



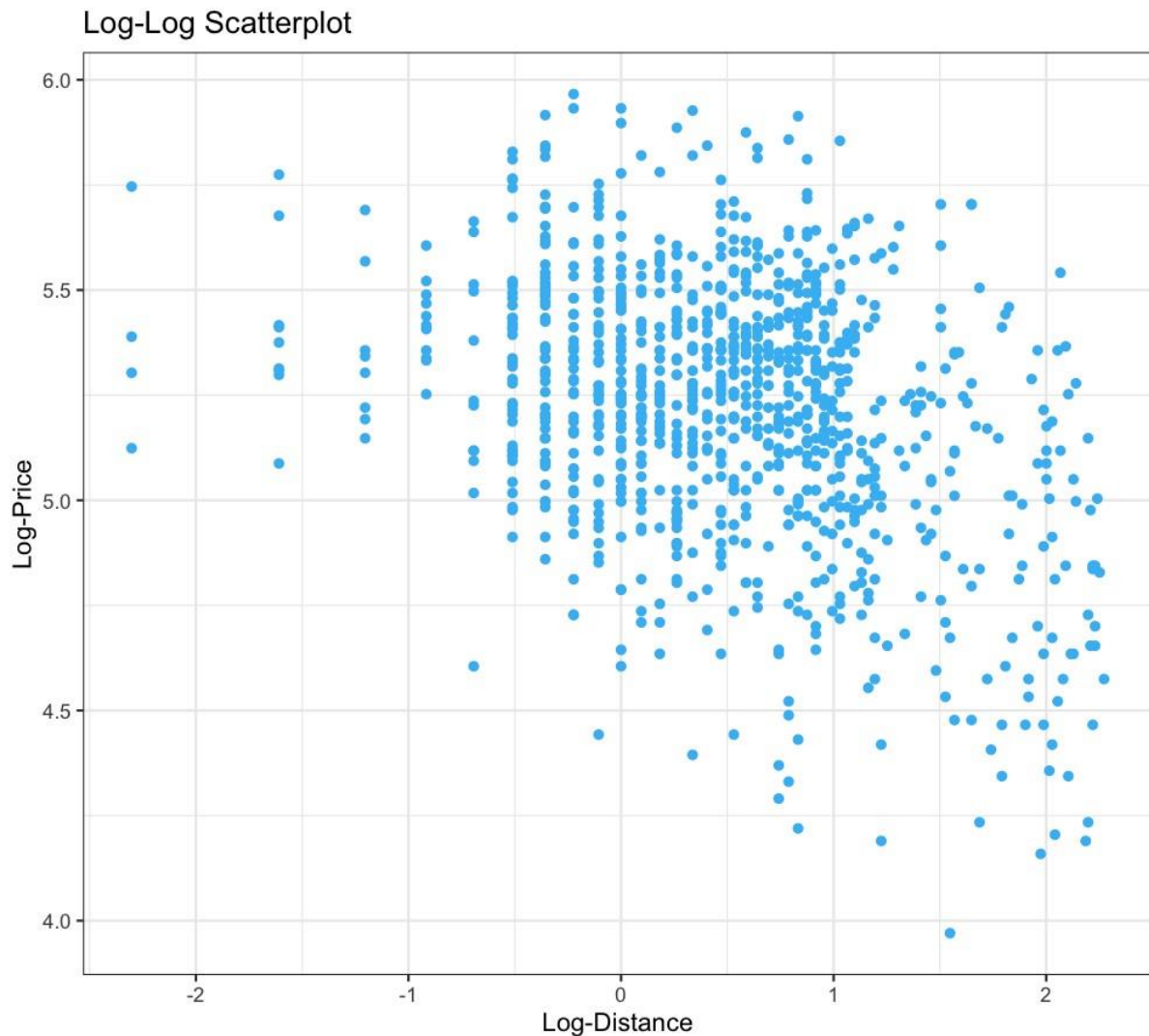
Although the pattern is not strictly linear, but we can see that as the log of distance increases, the price tends to decrease overall. However, the data is more scattered and does not follow a strong linear trend, suggests that the relationship might not be well captured by a linear model in this form. There are also outliers present, especially for short distances ($\text{log-distance} < 0$).

Since the log of distance is used, the relationship reflects proportional changes in distance rather than absolute changes. A linear regression model would suggest that absolute changes in price occur as the distance increases exponentially (as shown by the log transformation). This is helpful when smaller changes in distance close to the centre have larger effects on price than larger changes farther from the centre.



Compared to the level-log scatter plot, the points of the log-level plot are more clustered around smaller distances (0 to 2.5 km), and there's more variation at larger distances, but there is still an overall downward trend. Some outliers are still visible at both small and large distances, especially for hotels at greater distances that still have relatively higher prices. However, the log transformation of price has helped to reduce the impact of extreme price values, as seen in the more concentrated distribution of points.

In this specification, the log of price is used, meaning that percentage changes in price are associated with absolute changes in distance. This is useful to capture how much a fixed increase in distance (for example: 1 km) leads to percentage changes in price.



The points are more evenly distributed, and the relationship between the two variables is clearer, following a downward linear trend. In terms of outliers, the log-log transformation already compresses both the large price values and large distance values, which helps in reducing the effect of extreme observations.

The log-log model captures proportional changes in both price and distance. This means that the model explains how percentage changes in distance affect percentage changes in price. For example, a 10% increase in distance would lead to a certain percentage decrease in price. This model is particularly useful if the relationship between price and distance operates on relative changes rather than absolute changes.

Model choice

Among these three, the Level-Log specification fits my dataset best. As explained earlier about the Lowess model in section 2.4, the dataset has some fluctuations in price in the first 0-2km, where the curve initially shows a sharp decline in hotel prices. This means the closer the hotels are to the city centre, the more influenced the price is, in which prices being notably higher for hotels located closer to the city centre. And among the three models, the Level-Log addresses that most properly, with the interpretation of proportional changes in

distance result in absolute changes in price, which is most helpful when smaller changes in distance close to the centre have larger effects on price than larger changes farther from the centre like this dataset.

Although the Log-Log model has a clearer relationship between the two variables, its interpretation of percentage change relatively between the two variables is not effective for this dataset and its purpose of finding the best hotel deal. For example, to find a cheap relatively cheap hotel, it is difficult and not efficient to determine and understand how many percent of price is explained by a percentage change in distance. Likewise, while the LogLevel model has the preferable interpretation of absolute distance in miles, this model assumes that absolute changes in distance lead to percentage changes in price. However, this assumption is not appropriate when the relationship between distance and price is not fixed in this way, especially if changes in price don't consistently respond to changes in distance like this dataset.

c) Now, fit a **linear regression model** based on the specification you selected (log-log, loglevel, or level-log). The regression equation will depend on your chosen model. After estimating the model reproduce the estimated regression line on the scatterplot (just like in step 2.4). Present the regression results in a **table format**, discuss their statistical significance and interpret them. Note that interpretation of the regression coefficients will depend on the model that you have chosen.

Coefficient	Estimate	Std. Error	T value
Intercept	205.351	2.008	102.24
Slope	-23.895	2.239	-10.67

Multiple R squared	0.101
Adjusted R squared	0.1001

Intercept (205.351):

- The intercept represents the estimated hotel price when the log of distance is zero, which corresponds to a distance of 1 km/mile from the city center.
- With a value of 205.351, this means that the average price for hotels located at 1 km/mile from the city centre is approximately 205.351 units.
- The t-value for the intercept is 102.24, which is very large. This indicates that the estimate of the intercept is 102.24 standard deviations away from zero, meaning the average price of hotels located at 1km from the city centre is significantly different from zero.
- A t-value this large indicates that the intercept is highly reliable and has a strong influence on the model. The fact that it's so far from zero confirms that the estimated price for hotels located at 1 km is statistically sound and relevant.

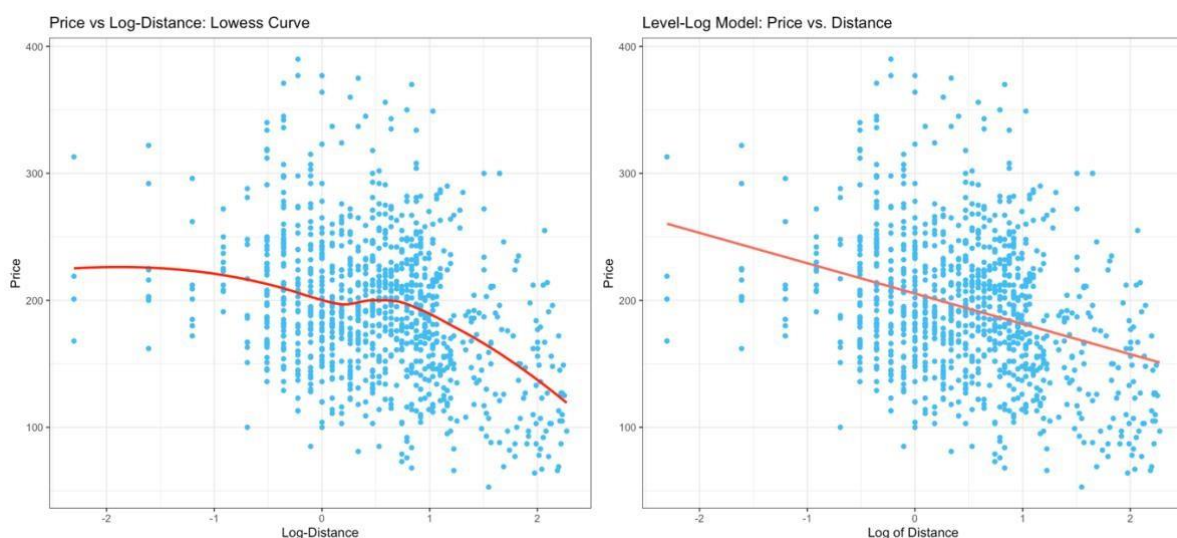
Slope for Distance (-23.895):

- The coefficient for log-distance is -23.895, meaning that for every 1% increase in distance from the city center, the hotel price decreases by 23.895 units

- This negative relationship supports the expectation that hotels farther from the city center tend to have lower prices, and the log-distance transformation means that proportional changes in distance lead to absolute changes in price.
- The t-value for the log-distance coefficient is -10.67, which is very large. This indicates that the coefficient estimate of -23.895 is 10.67 standard deviations away from zero, confirming that the effect of distance on price is statistically significant. The negative sign indicates that as distance increases, price decreases.
- A large t-value with this magnitude suggests that log-distance is a strong predictor of hotel prices. This relationship is not likely due to random chance, but reflects a meaningful association between distance and price.
- Typically, a t-value above 2 or below -2 is considered statistically significant, but a tvalue of -10.67 offers strong evidence that distance plays a critical role in determining hotel prices.

R-squared (0.101): The R-squared value is 0.101, indicating that the model explains approximately 10.1% of the variance in hotel prices based on distance. While this shows that distance has some predictive power in determining hotel prices, a large portion of the price variability is explained by other factors not captured by this model. These could include elements like proximity to airports, major landmarks, or hotel star ratings—factors that we have chosen to leave out of this model to focus solely on the relationship between price and distance

Below are the Lowess model and the estimated regression line on the scatterplot



2.6 Linear piecewise spline

Coefficient	Estimate	Std. Error	T value
-------------	----------	------------	---------

Intercept	231.99	9.5287	24.347
lspline 1	-28.0978	10.4816	-2.681
Lspline 2	-10.0804	0.9937	-10.145

Multiple R squared	0.1227
Adjusted R squared	0.1209

Intercept (231.99):

- The intercept represents the predicted hotel price when the distance (log-spline) is zero. In practical terms, this means the price of a hotel at the knot point or distance where the model's slope changes.
- With an estimate of 231.99, it suggests that the average hotel price at this point is around 231.99 units (such as dollars or euros).
- The t-value for the intercept is 24.347, which is very large. This indicates that the estimate of the intercept is 24.35 standard deviations away from zero. Such a high t-value means the intercept is statistically significant, and the model confidently estimates the baseline price at this specific distance.

Spline 1 Coefficient (-28.0978):

- The first spline coefficient of -28.0978 represents the rate at which prices decrease as distance increases up to the knot point. In this case, for every unit change in distance up to the first knot, hotel prices decrease by approximately 28.10 units.
- The t-value for spline 1 is -2.681, which indicates the coefficient is 2.68 standard deviations away from zero. This suggests that the decrease in price up to the first knot is statistically significant, although not as strongly as the other coefficients.

Spline 2 Coefficient (-10.0804):

- The second spline coefficient of -10.0804 represents the rate at which hotel prices decrease after the knot point. Here, for every unit change in distance beyond the knot, hotel prices decrease by approximately 10.08 units.
- The t-value for spline 2 is -10.145, which is quite large, indicating that the price decrease beyond the knot is highly significant and the relationship is reliable. The fact that this t-value is high suggests that the model is very confident about the effect of distance on price after the knot.

Statistical Significance:

- We can see that the t-values across all coefficients are larger than 2 in magnitude, meaning that these coefficients are statistically significant at conventional levels (usually above 2 or below -2 is significant).

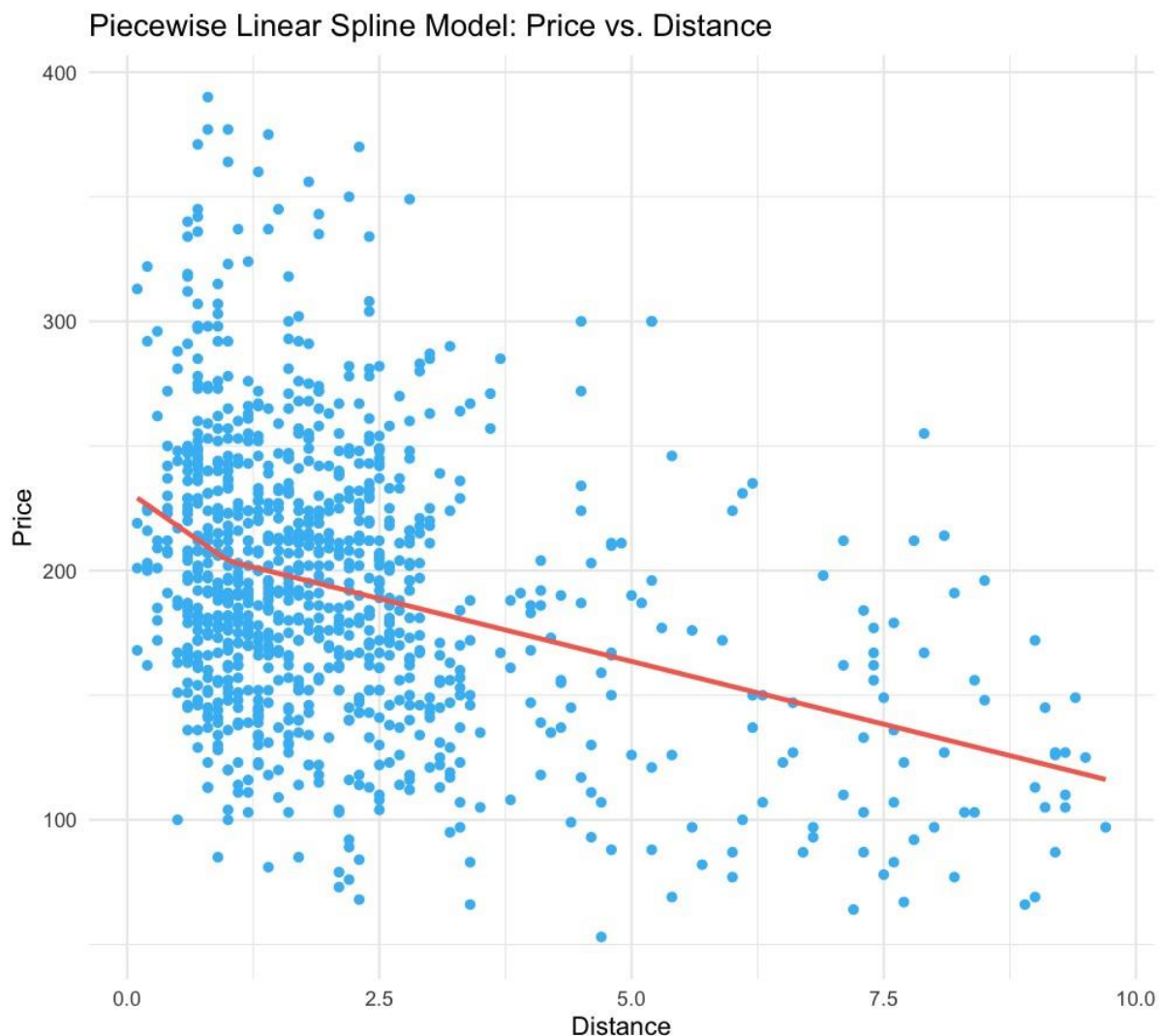
R-squared:

- The Multiple R-squared of 0.1227 indicates that 12.27% of the variance in hotel prices is explained by the model. This is a modest level of explanatory power, suggesting that other factors not included in the model (e.g., hotel amenities, star rating, proximity to landmarks) account for much of the variation in price.

Model Coefficients Interpretation:

The model suggests that hotel prices decrease significantly as distance from the city centre increases, with the decline being steeper up to the knot point (where the first spline coefficient applies). After the knot point, the rate of price decrease becomes less steep but still continues, as indicated by the second spline coefficient of -10.08. In general, this means that proximity to the city centre has a large impact on hotel prices, but the impact diminishes slightly after a certain distance.

Below is the Piecewise Linear Spline model with the knot point of 1.



2.7 Choosing the best model

To effectively choose which model is best, the purpose of the analysis should be reiterated. As mentioned above, the main goal of the analysis is to capture the general trend between price and distance patterns of hotel in Paris, that's why we chose to drop all extreme values of price and distance by a certain threshold, holding other factors constant (star ratings and geological reasons). Therefore, to choose among the three models, I will choose the model that explains the dataset best and align with the initial purpose of the analysis.

First, I compare the three models

	(1)	(2)	(3)
(Intercept)	216.740 (2.578)	205.351 (2.008)	231.991 (9.529)
distance	-10.739 (0.912)		
log_distance		-23.895 (2.239)	
lspline(distance, knots = cutoff_point)1			-28.098 (10.482)
lspline(distance, knots = cutoff_point)2			-10.080 (0.994)
Num.Obs.	1016	1016	1016
R2	0.120	0.101	0.123

(1), (2), and (3) are Linear model, Level-Log model and Spline model respectively. As their coefficient and r-squared values are already interpreted above, the below comparison will be a general analysis on the three models' strengths and weaknesses, statistically and practically, from which, we can choose the best model that fits the dataset.

Model (1): Linear Regression Model

Strengths:

- The model is easy to interpret and implement. It provides a straightforward relationship between distance and price, where every unit increase in distance leads to a fixed decrease in price.
- The coefficient for distance is directly interpretable in terms of how many units that hotel prices decrease per kilometre/mile of distance from the city centre.

Weaknesses:

- The model assumes that the relationship between distance and price is linear, which might not reflect the real-world situation in a city like Paris. Hotels close to the city centre might have very different pricing dynamics than those farther away.
- With an R-squared of 0.120, this model explains a relatively small percentage of the variance in hotel prices compared to the spline model. This indicates that important price variations due to other besides price, are not captured.
- The model does not capture potential non-linear relationships, such as different rates of price decline at various distance ranges, which are likely present in a city with a complex layout like Paris.

Model (2): Log-Level Model

Strengths:

- By using the logarithm of distance, the model captures the proportional effect of distance on price. This is useful when small percentage changes in distance (such as moving from 1 km to 2 km) have a larger effect than larger changes (moving from 10 km to 11 km).
- The log transformation makes it suitable for cases where price changes rapidly over short distances (closer to the city centre), which is a realistic assumption for Paris.
- The log transformation reflects diminishing returns — meaning that further away distances have a smaller effect on price, which is more realistic for cities where central areas are more expensive and price drops level off as distance increases.

Weaknesses:

- This model has a low r-squared. With an R-squared of 0.101, this model explains the least amount of variance in hotel prices among the three models. This suggests that other factors besides distance may explain a large part of the price variation.
- While useful for proportional relationships, interpreting the coefficients can be more challenging compared to a simple linear model, as it's based on percentage changes rather than direct unit changes in distance.
- The log transformation tends to flatten out at higher distances, which might oversimplify price variations for hotels located far from the city centre.

Model (3): Spline Model (Piecewise Linear Model)

Strengths:

- The spline model allows for a non-linear relationship between price and distance, capturing different rates of change at different ranges of distance. This is more realistic for cities like Paris, where proximity to landmarks can have a significant impact on price early on, but the effect diminishes at greater distances.

- By using splines, the model can capture sharp changes in price at certain distances (before and after the knot), allowing for multiple pricing trends within the same model.
- With an R-squared of 0.123, the spline model explains more variance in hotel prices than the other two models, suggesting that it fits the data better. It is especially useful in a city like Paris where pricing dynamics vary significantly depending on location.
- The model reflects the reality that hotel prices in Paris decrease sharply as you move away from the centre initially but level out or decrease more slowly at greater distances.

Weaknesses:

- The spline model is more complex to interpret and requires choosing appropriate knots. It's not as straightforward as the linear model in terms of interpretation.
- Because of its flexibility, there's a risk of overfitting the data, especially if too many knots are introduced. This might result in the model being too tailored to the current dataset and less generalizable to new data.
- While the flexibility of the spline model offers better performance, it can be harder to explain to non-technical stakeholders who might prefer a simpler linear model.

Among the three models, Piecewise Linear Spline model is the optimal option to choose. Firstly, its statistical figures are the most significant, with the R-squared of 0.123, which is the highest among the three models. This suggests that the model not only fits the data best, but also have the most potential predictive power among the three for similar cases. Secondly, the models align with the initial purpose of the analysis. When customer look at the analysis, they would want to get the best hotel deals. The difference of price variation at the centre area and the others will be considered by customers when making decisions about tourism planning and hotel selection, and this model provides the best fitting visualisation to the real-world scenario, where prices of hotels near the city centre will experience a steep price premium, and even more on the weekend, when the data is observed; compared to those located just a little farther away, which might offer a much better value. This model also explains the diminishing effect of hotel price well, which is missing from the other two models.

The spline model also covers the airport and suburban effect, in which beyond a certain distance (e.g., 7-10 km), hotel pricing is less about proximity to Paris landmarks and more about proximity to transportation hubs like airports or suburban attractions. The second slope in the spline model captures this reality, as pricing patterns shift from inner-city dynamics to more suburban or transport-related factors. Furthermore, real estate developers and urban planners can use the spline model to understand the demand elasticity for hotels at various distances from the city centre. This is particularly relevant for planning new hotels, deciding on property values, or setting zoning policies. The model identifies where pricing plateaus and where it is worth investing in hotels closer to or farther from the city centre.

One more thing about the spline model that makes it stand out between the three models is that it is more customizable to a city like Paris, where pricing could be influenced by specific districts. Thanks to the ability to introduce multiple knots if necessary, the model can capture

subtle shifts in pricing across different neighbourhoods, making it better tailored to the data compared to rigid linear or log-level models.

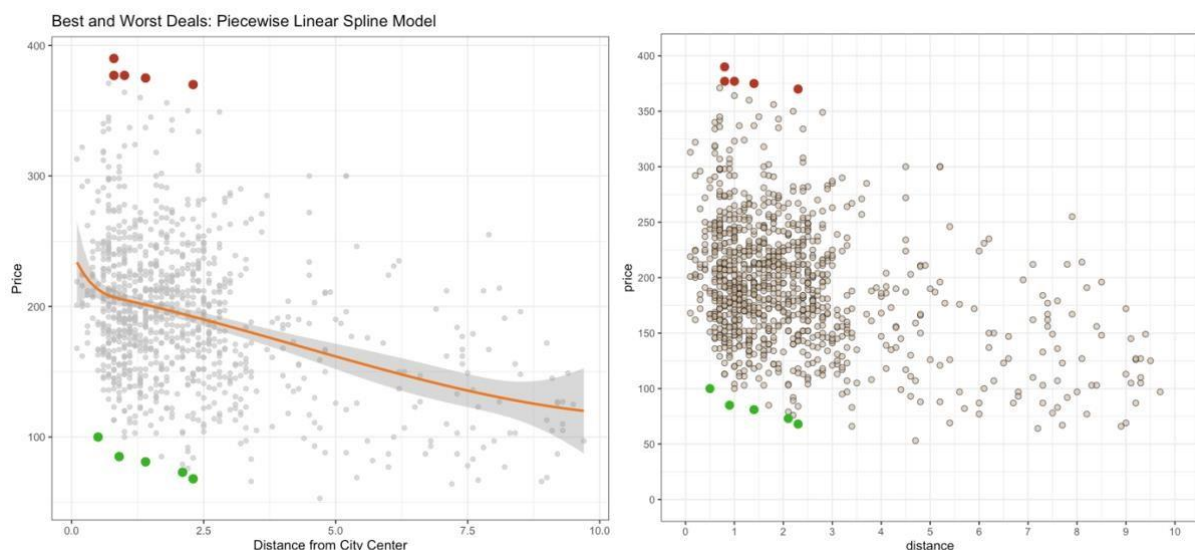
After choosing the best model – piecewise linear spline model, we can compute the residuals table, which can show us the best deal and the worst deal among all the hotels in Paris.

	Hotel_id	Price	Residuals
Best deal	12451	68	-122.79
Worst deal	13399	390	180.49

The best deal among the listed hotels is hotel 12451, as it has the most negative residual (122.7882). A more negative residual indicates that the hotel is priced significantly lower than what the model predicts. In other words, hotel 12451 offers a price that is much lower than expected based on the general price-distance trend.

The worst deal among the listed hotels is hotel 13399, as it has the most positive residual (180.4877). Conversely, a higher positive residual indicates that the hotel is priced significantly higher than what the model predicts. In other words, hotel 13399 is charging much more than expected based on the general price-distance trend.

Below is the scatterplot of data showing the estimated regression line for the best model and five best deals (in green) and five worst deals (in red)



Both scatter plots show the 5 best deals and five worst deals. While for the left plot, I include the confidence interval for the predicted values of the model. Specifically, this is typically a 95% confidence interval, meaning that there is a 95% probability that the true regression line for the relationship between price and distance falls within this shaded region.

Result

Through the analysis, customers can grasp valuable insights into the relationship between hotel prices and their distance from the city centre in Paris. By examining the validity of three models, we were able to choose the piecewise linear spline model, which helps to capture the

non-linear nature of the relationship between the two variables of interest. This non-linear nature also reflects the complexities of pricing patterns in a major city like Paris, suggests that this model can be used in similar real-world scenarios.

The model shows that hotel prices decline sharply with distance from the city centre, but the rate of this decline slows down at greater distances, known as diminishing effect. This nonlinear pattern is realistic, as proximity to key landmarks (such as famous tourist attractions) or central areas can significantly impact hotel pricing, while hotels farther from these areas experience a more gradual reduction in price.

Finally, the identification of the best and worst hotel deals through residual analysis highlighted outliers in the data. Hotels that are underpriced relative to their location offer great value to potential travelers, while overpriced hotels indicate poor value. Therefore, this analysis can be extremely useful for both consumers looking for affordable accommodations and hotel owners assessing their pricing strategies.

Its limitation

- All of the suggested model just focuses on the relationship between price and distance variable, while omitting other potential factors such as star ratings, room amenities, proximity to specific landmarks or transportation hubs or seasonality.
- Even though the spline model has the highest among the three models, this r-square value is still small, which indicates that distance alone can just explain a small portion of the variation in the hotel prices.
- While the piecewise spline model captures non-linearity, there may still be other forms of non-linearities that a more complex model could reveal. The piecewise model may not fully capture the nuanced effects of distance, especially in relation to key areas within the city.
- The dataset used only focuses on a specific time and place, which means any conclusions drawn from it might not be generalizable to other cities or even other seasons in Paris.

To better model the relationship between price and distance, or to better explain the dataset in general, these suggestions below can be beneficial:

- In the future, I would use models that could include variables such as hotel star ratings, room size, availability of amenities, and seasonality to better explain the variability in hotel prices. These factors likely play a significant role in determining price and would enhance the model's predictive power.
- In addition to the piecewise spline model, other non-linear models such as quadratic regression or log-transformed models could be utilised. These models may better capture subtle variations in how distance affects prices, particularly at the extremes of the distance range.
- Fine-tuning the cutoff point to match the specific data and characteristics of each city can help capture a more accurate price-distance relationship. Since every city has its own unique pricing patterns—shaped by factors like tourism demand, transportation access, and local economic conditions—applying a one-size-fits-all approach may not

work. By adjusting these thresholds based on the city and dataset, we can create a more precise and meaningful analysis that truly reflects real-world trends.

- Rather than dropping all extreme values, more sophisticated techniques could be applied to identify outlier clusters or segment the hotel market into distinct groups. For instance, high-end hotels might follow different pricing rules than budget hotels,

and modelling these separately could provide more accurate predictions. This will require us to use other complex models that can capture the non-linear nature of this kind of dataset