

HOTEL PRICING DYNAMICS ANALYSIS

Applying R to Analyse Pricing Trends and Enhance Value for Travelers

Prepared by: Hailey Do

Date: 28 December 2024

TABLE OF CONTENTS

1. Explanation	3
2. Distribution of the Variables	5
3. Addressing Extreme Values	7
4. Linear regression.....	11
5. Log transformations.....	13
6. Linear piecewise spline	19
7. Choosing the best model	22

1. Explanation

Filter the rows in the dataframe df where the city is Rome, year is 2017, in November (11) and weekend is 0 (indicating weekdays only)

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

Separate the values in the accommodation type column into 2 parts using the '@' as a separator. The first part is stored in a column named 'garbage' and the second part in a column named 'acc_type'.

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

Separate the distance column values into 2 parts using the space (' ') as a separator. The first part is stored in a 'distance' column, and the second part ('mile') is stored in a column named 'miles'.

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

Remove the 'garbage' and 'miles' columns from the dataset (unnecessary columns)

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

Filters the df to only include the rows where the acc_type is 'Hotel'

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

Select only the 'hotel_id', 'distance', 'price', 'neighbourhood' and 'starrating' columns.

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





Convert the distance column from the character format to a numeric format so we can do calculations or comparisons on it.

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

Remove any duplicated rows based on all the selected columns in the df.

```
distinct()
```

The command `datasummary_skim(city_data)` generates a table overview of the dataset which includes unique values, mean, standard deviation, min, median, max and histogram.

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	Histogram
hotel_id	734	0	17270.7	1269.8	15186.0	17264.0	20045.0	
distance	84	0	2.2	3.1	0.1	1.1	16.0	
price	191	0	109.7	83.7	32.0	86.5	1145.0	
starrating	6	0	3.2	0.9	1.0	3.0	5.0	

Discussions:

There are no missing data points in the dataset, which is a strong indicator of the dataset's completeness and reliability. The dataset contains 734 unique hotels, with a wide range of distances, prices, and star ratings. This variety allows for a detailed analysis of how different factors like proximity to the city centre, price, and hotel ratings have any interactions with each other's.

Distance:

The average distance of hotels from the city centre is 2.2 miles, with a range from 0.1 to 16 miles. The right-skewed distribution indicates that while most hotels are close to the city centre, a smaller number are located further away. This skewed distribution could be attributed to the higher demand for centrally located accommodations, where most tourist attractions are typically concentrated. The few hotels located further from the centre might cater to a niche market, such as budget-conscious travellers or those seeking quieter locations with high-luxury hotel demanded.

Price:

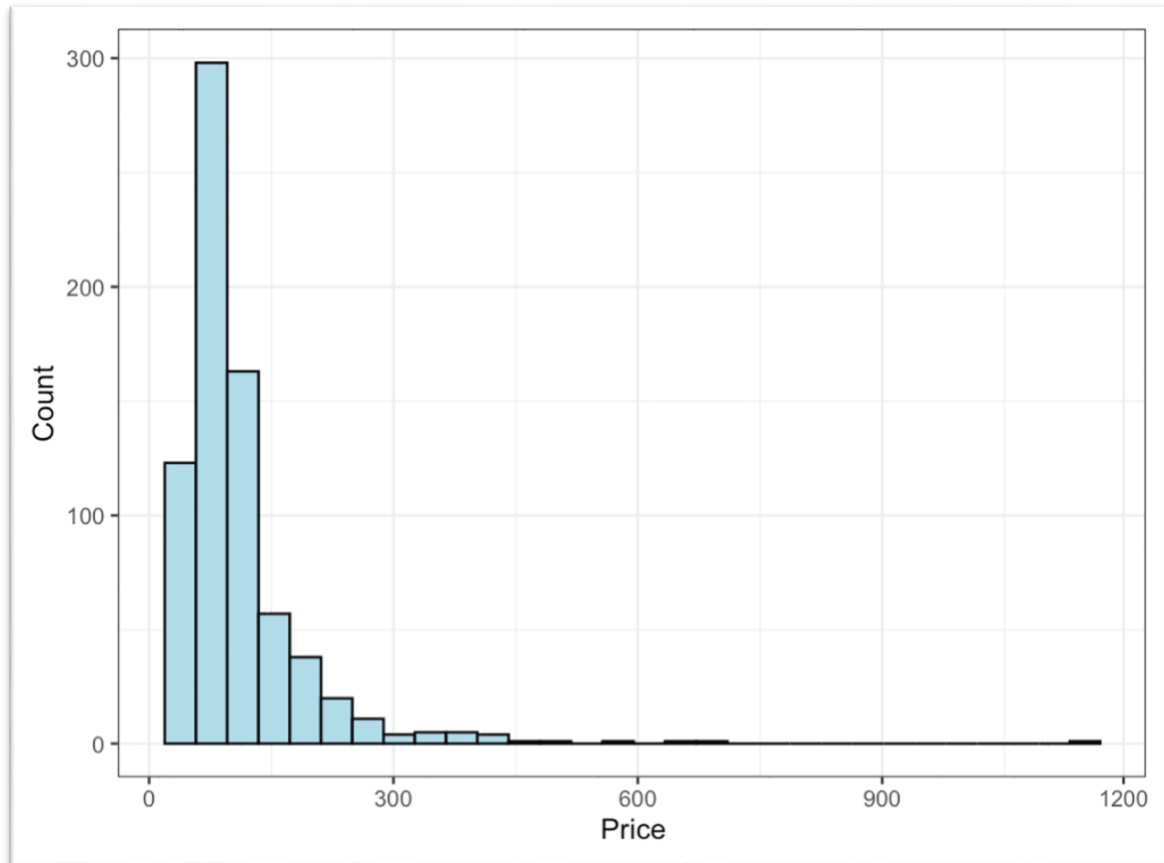
Similarly, the average hotel price is \$109.7, and the distribution is highly skewed, ranging from \$32 to \$1,145. The presence of high-priced outliers reflects the wide disparity in hotel prices, possibly due to differences in hotel quality, amenities, or proximity to popular attractions. The skewness also suggests that many of the hotels fall into the lower price range, likely making them more accessible to budget-conscious travellers, while a few luxury hotels significantly charge higher price than typical pattern of hotels.

Star Rating:

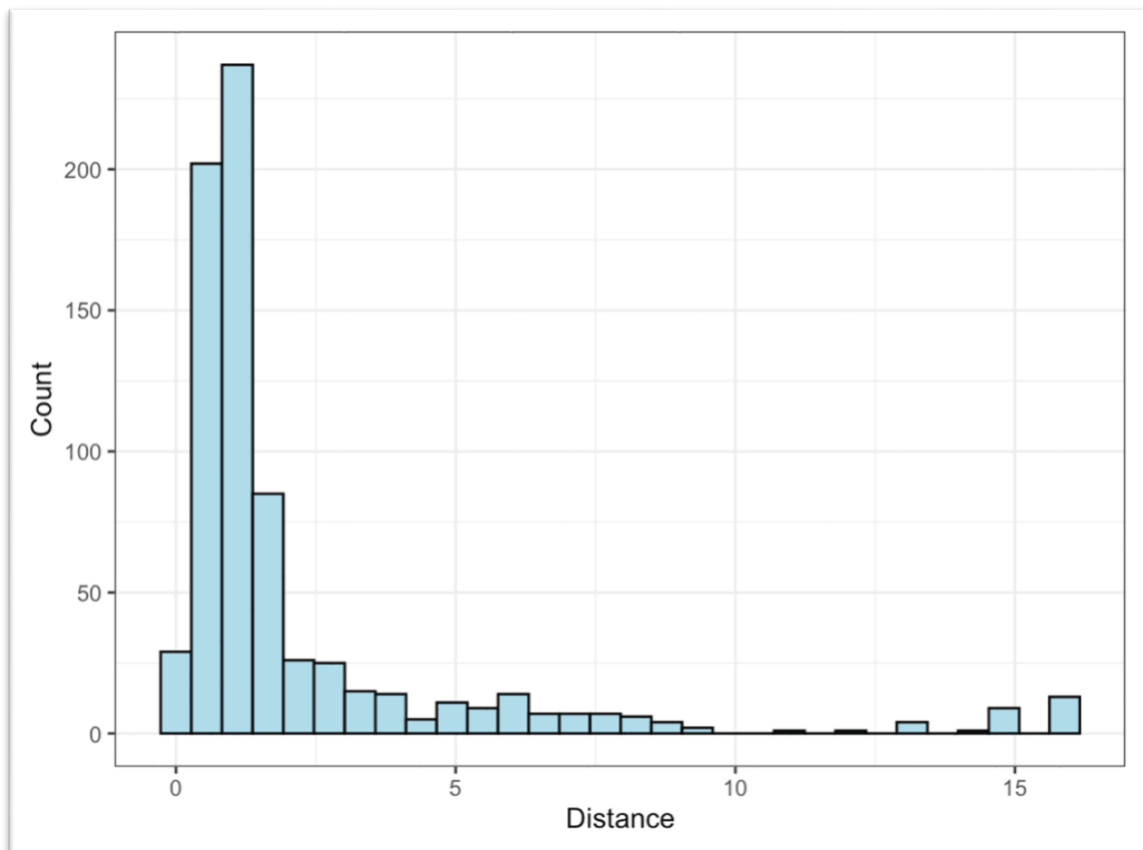
The star ratings of the hotels are near-normally distributed, with an average of 3.2 stars, indicating that most hotels are mid-range. This spread suggests that travellers have a variety of options, from reasonable budget accommodations to more luxury hotels, although those with higher star ratings are less common.

2. Distribution of the Variables

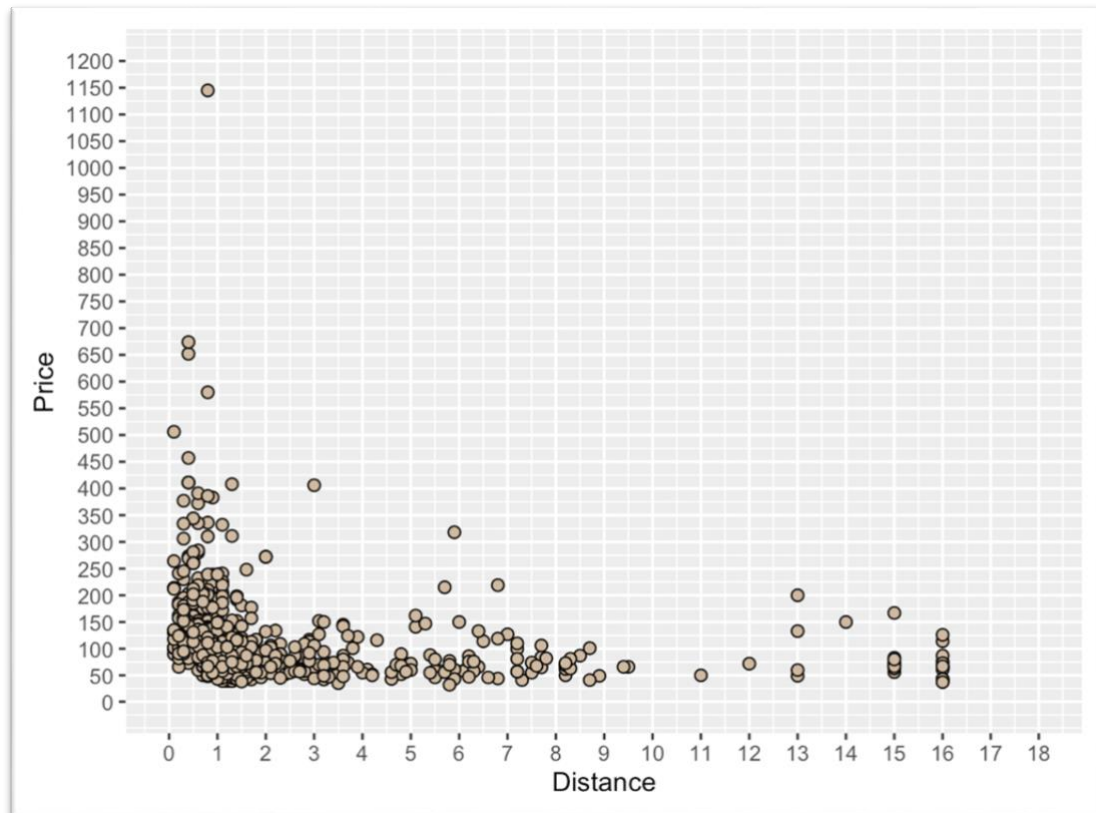
The histogram for the distance (distance from the city centre):



The histogram for the price (hotel price):



Scatterplot with distance on the x-axis and price on the y-axis to visualise the relationship between these two variables.



i/

The distribution of hotel prices and distance are both right skewed, meaning most hotels have a reasonable price, with a few hotels having significantly higher prices. Most hotels are being close to the city centre (around 0 - 2 miles), and a few hotels are farther (up to 16 miles).

ii/

There are noticeable outliers in the histogram of the price, particularly in the higher price range (above \$600), which suggests some luxury hotels or special cases that charge significantly more than the average.

There are some hotels located far from the city centre (above 10 miles). These are clear outliers, as most hotels are clustered around to the city centre.

iii/

The scatterplot indicated a **negative relationship** between 2 variables because when the distance from the city centre increases, the price of hotels tends to decrease.

There are several outliers in the scatterplot, particularly some higher-priced hotels (above \$600) located at short distances (close to the centre = 0 miles)

A clear cluster of hotels is observed between the price range of \$50 and \$300, particularly within 0-5 miles from the centre. This indicates that most hotels are priced within this range and located closer to the central city areas. Moreover, there is a decline in price concentration beyond 5 miles, with prices generally dropping but a few remaining high-priced outliers beyond 10 miles.

3. Addressing Extreme Values

a/

Firstly, I decided to exclude the 5-star hotel because of some main reasons:

- **Reason:** I decided to exclude 5-star hotels because they often do not follow the same price-distance relationship as mid-range hotels (3 to 4 stars).
- **Key Point: Mid-range hotels** (3 to 4 stars) are more representative of typical travellers and follow a clearer trend where **price decreases with distance** from the city centre. Since the **median star rating is 3**, focusing on this range makes the analysis more representative and avoids skewing by luxury hotels.

Secondly, I handle with **Price** and **Distance outliers** by conducting evaluations for them with the interpretations below:

```
# Evaluate the data of distance > 10 miles and price > 500
ev_distance <- filter(city_data, distance > 10)
ev_price <- filter(city_data, price > 500)
```

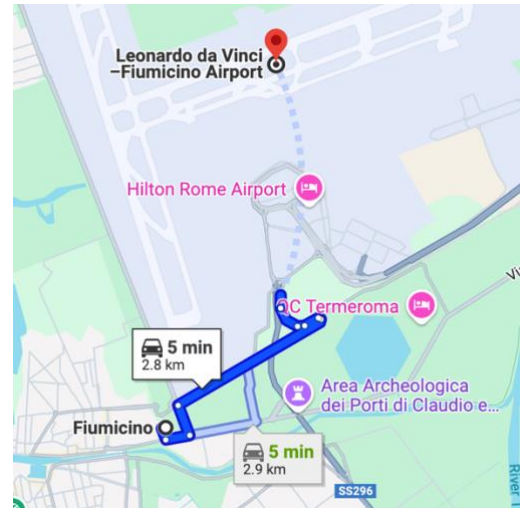
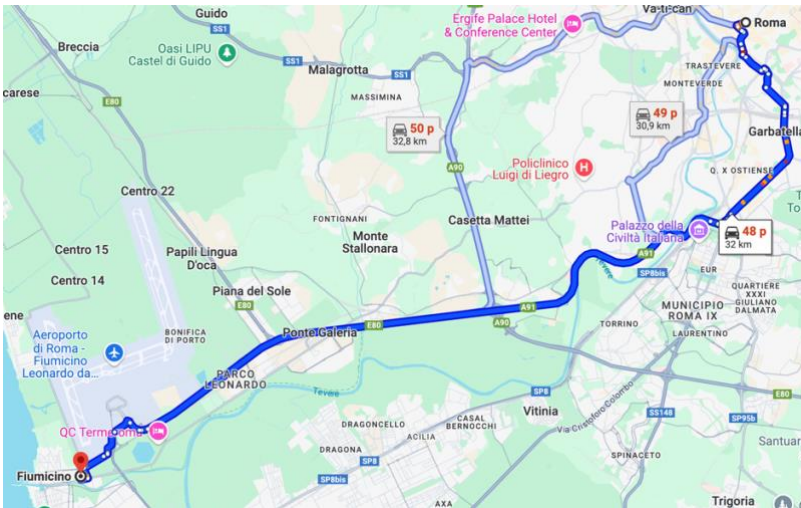
- **Distance Outliers (distance > 10 miles)** these are outliers that do not align with the typical travel pattern and may not follow the usual **price-distance relationship**.
- **Price Outliers (prices > \$500):** While high prices normally can provide insights but for this analysis (we are focusing on price vs. distance relationship only), I will drop outliers where **low or high star ratings (2-star and 5-star)** in the price outliers data **make price comparisons less relevant**.

Rationale: Dropping extreme values for price and distance ensures that the analysis focuses on **mid-range hotels** and reflects the expected relationship between **price and distance**.

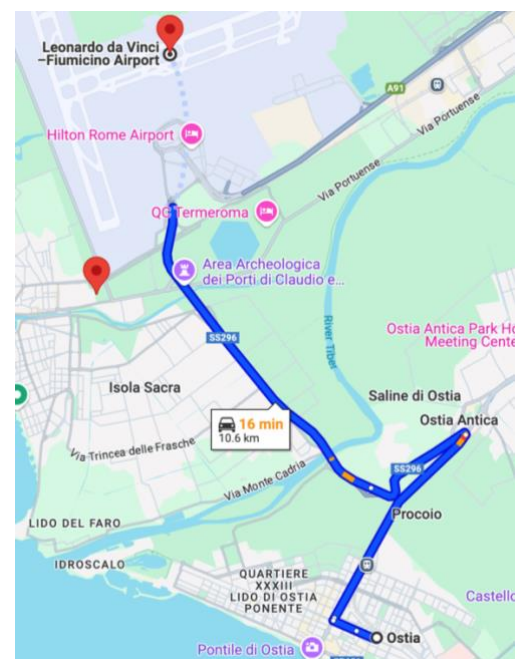
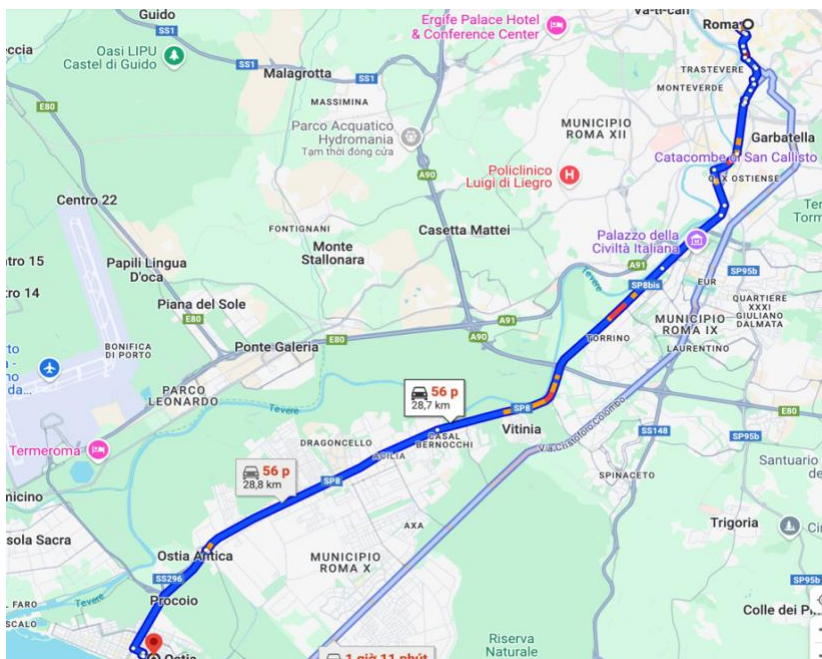
In the evaluation for the price variable (> \$500), I decided to drop these outliers with 2-star and 5-star ratings as they do not offer relevant insights for analysing the price-distance relationship. In other word, I drop the extreme values in **price variable**.

	hotel_id	distance	price	neighbourhood	starrating
1	16676	0.8	1145	Repubblica	5
2	18681	0.4	674	Spanish Steps	5
3	15973	0.4	652	Ludovisi	5
4	16601	0.8	580	Repubblica	2
5	18784	0.1	506	Spanish Steps	5

When it comes to the **distance variable**, I decided to remove some hotels beyond **10 miles**. Even though when I consider with other variables: starrating; some of them have the median starrating at 3 stars which is fairly medium type of hotel, it would not be the potential options for customer, who likely to choose the ones near the city centres with multiple activities rather than only near the airport. Specifically, **Fiumicino** and **Ostia** are close to the Leonardo da Vinci-Fiumicino airport, which are examining as far away from the city centre (not the one that we are looking for), so I will **drop** those extreme values in the distance variable.



Fiumicino neighbourhood (regarding distance to the city centres and the airport)



Ostia neighbourhood (regarding distance to the city centres and the airport)

The distance variable evaluation result is shown below:

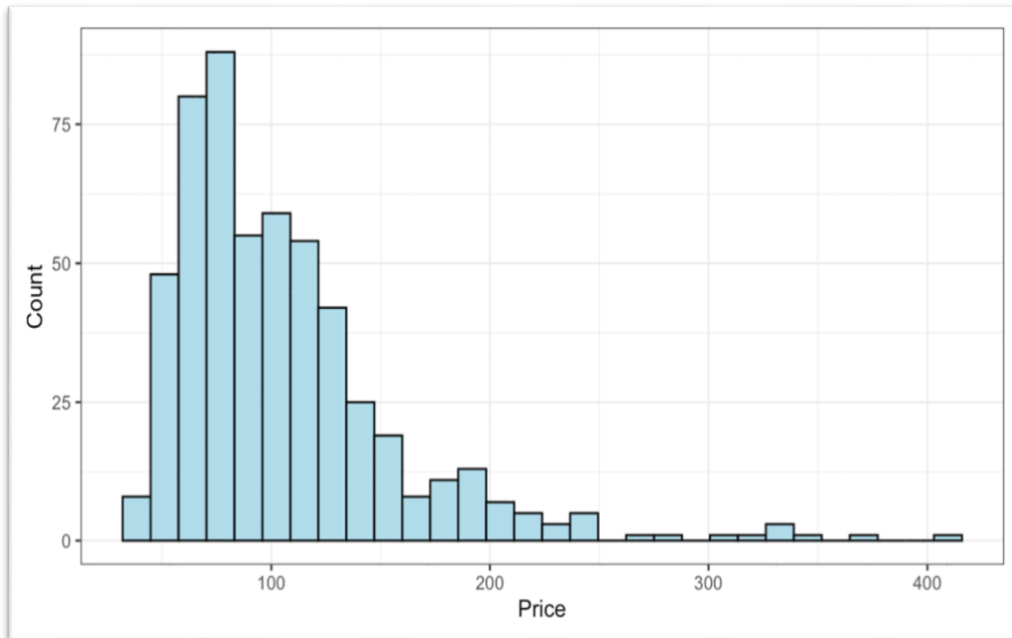
	hotel_id	distance	price	neighbourhood	starrating						
1	15810	15	56	Fiumicino	2	17	15812	14	150	Fiumicino	4
2	15859	16	65	Fiumicino	2	18	15813	15	83	Fiumicino	4
3	15863	15	67	Fiumicino	2	19	15816	16	114	Fiumicino	4
4	16384	16	38	Ostia	2	20	15830	15	75	Fiumicino	4
5	15822	16	66	Fiumicino	3	21	15832	15	167	Fiumicino	4
6	15824	16	76	Fiumicino	3	22	15840	16	87	Fiumicino	4
7	15833	15	73	Fiumicino	3	23	15841	13	133	Fiumicino	4
8	15836	15	69	Fiumicino	3	24	15846	15	63	Fiumicino	4
9	15865	16	73	Fiumicino	3	25	15856	16	126	Fiumicino	4
10	15866	15	80	Fiumicino	3	26	16383	16	67	Ostia	4
11	15868	16	61	Fiumicino	3	27	16394	11	50	Ostia	4
12	16386	16	43	Ostia	3	28	16404	12	72	Ostia	4
13	16387	16	43	Ostia	3	29	15828	13	200	Fiumicino	5
14	16388	16	37	Ostia	3						
15	16397	13	49	Ostia	3						
16	16408	13	60	Ostia	3						

b/

In other words, I will restrict the sample to with the data points of the **3-star, 3.5-star and 4-star hotels** (excluding too low star rating or too high star rating) using *filter* command and excluding the extreme values (in price and distance variables) as well. After restring the data, I got **540** observations left.

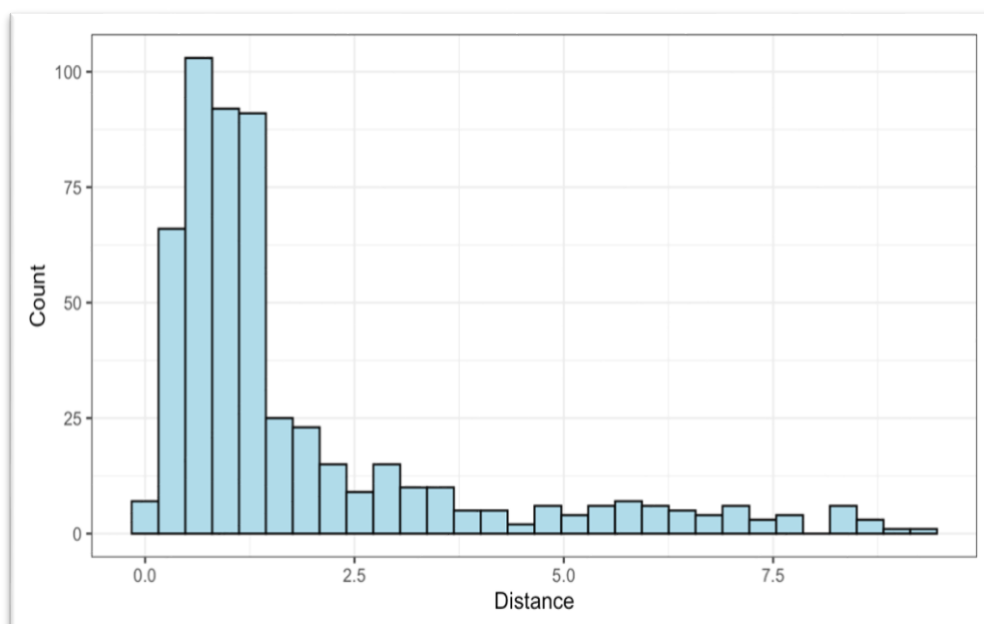
```
# Restrict the sample
city_data <- filter(city_data, distance <= 10 & starrating %in% c(3, 3.5, 4))
```

The histogram for **price** variable after conducting an evaluation:



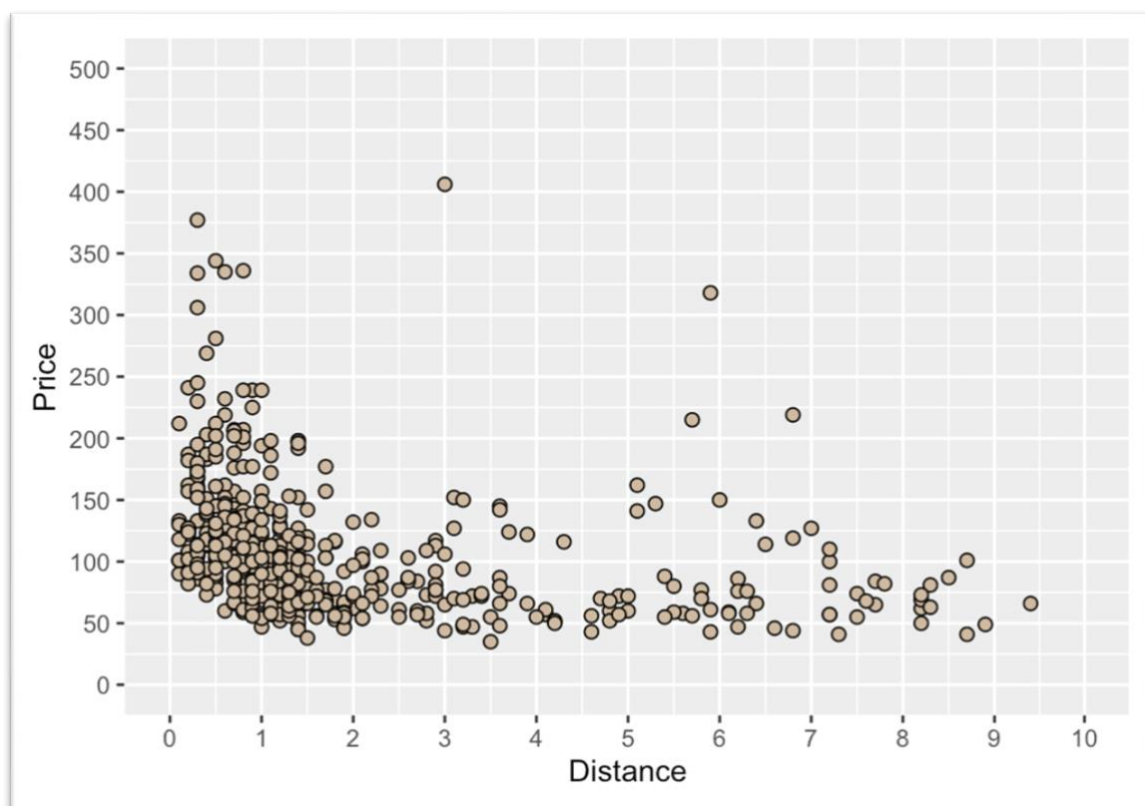
The histogram displays a right-skewed distribution of hotel prices, with most hotels priced between \$0 and \$100. The peak in this range highlights that cheaper hotels dominate the dataset. Compared to the original histogram, this refined evaluation reduces the long right tail by cutting off prices above \$500, which are considered outliers associated with luxury or high-end accommodations. This adjustment results in a more homogeneous distribution and helps to mitigate the influence of extreme values, making the dataset more representative of typical hotel prices.

The histogram for **distance** variable after conducting an evaluation:



The histogram for the distance variable shows a right-skewed distribution, with most hotels concentrated within 1.5 km of the city centre. As the distance increases, the number of hotels decreases sharply, with very few located beyond 2.5 km. This suggests that most hotels are centrally located, with a small portion of hotels extending up to around 7.5 km from the city centre. Compared to the original histogram, which had a longer right tail extending to 18 km, this refined version focuses on a more condensed range of typical data. The evaluation now emphasises 3, 3.5, 4-star hotels within 10 miles for further investigation.

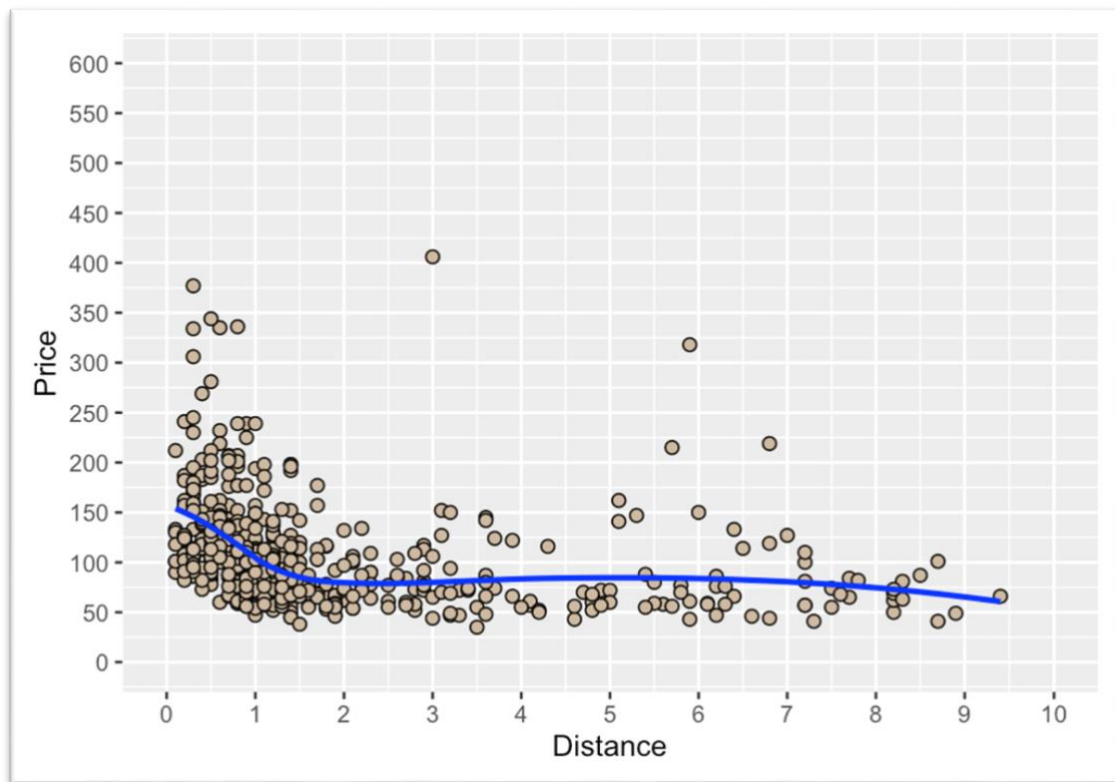
The scatterplot illustrates the relationship between hotel price and distance from the city centre, showing a clear negative trend. Prices tend to decrease as the distance from the city centre increases. Hotels closer to the centre (within 0 to 2 km) generally have higher prices, while those further away (beyond 5 km) tend to be lower-priced. This is somewhat counterintuitive, as one might expect hotels further from the centre to charge higher prices due to factors such as exclusivity or larger spaces. However, the data suggests that proximity to the city centre is a key driver of hotel pricing, reinforcing the idea that central locations are highly valued. Overall, this relationship highlights the importance of location in pricing strategies for hotels.



I might have one suggestion is to investigate if other factors such as star ratings, or seasonal demand could also play a role in determining prices. This could provide a more comprehensive understanding of hotel pricing dynamics in relation to distance. Additionally, exploring whether clusters in certain distance ranges represent specific hotel categories could offer deeper insights.

4. Linear regression

a/ Non-Parametric Lowess Fit



The Lowess curve shows a **non-linear relationship** between price and distance, especially for distances less than 2 miles. In the first 2 miles, the price drops significantly, indicating that hotels closer to the city centre are generally more expensive. Beyond 2 miles, the trend keeps remain, showing the price becomes more stable and does not decrease further with increasing price.

In other words, after around 2 miles, the price trend **plateaus**, showing that the hotels located further away from the city centre tend to have similar prices. The prices are stable at a certain level, suggesting that distance does not impact price significantly once hotels are father away. No sharp increases or fluctuations were observed beyond this distance, which further supports the observation of price stability after a certain distance.

Based on the shape of the Lowess curve, a linear model might not appropriate, especially for distances < 2 miles, where the drop in price is quite sharp. While a linear model might not capture the curved, but non-linear behaviour seen in the first few data points in distance. Besides, for distances > 2 miles, the relationship between price and distance become more stable, and a linear model could potentially be used to this range only.

b/ Fitting a linear regression model

$$Price = \alpha + \beta * Distance + \epsilon$$

The equation in the form:

The regression result in a table format using `summary()` command:

Regression 1 table format:

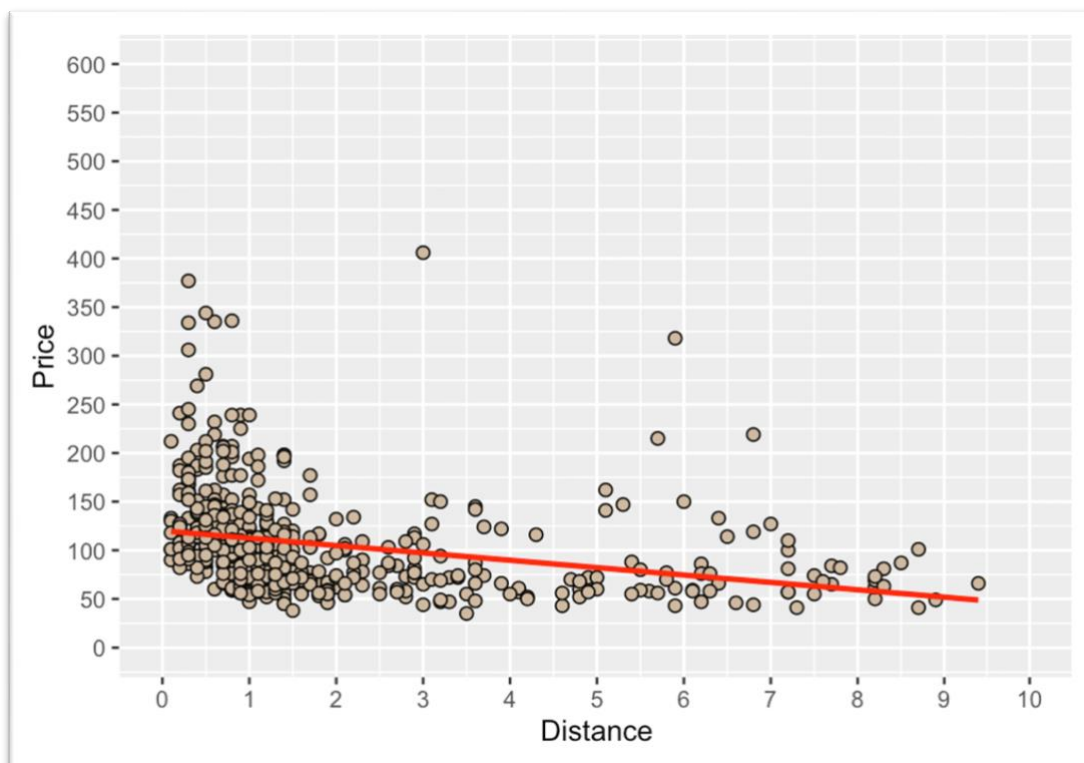
Variable	Coefficient	Std. Error	t-value	p-value
Intercept	120.063	2.989	40.171	<2e-16
distance	-7.583	1.111	-6.827	2.35E-11
	R-squared	0.07972		
	Adj R squared	0.07801		

The **intercept** 120.063 represents the predicted hotel price when the distance from the city centre is 0. This implies that at the centre of the city, the average hotel price is approximately \$120.063. This value is statistically significant with a very low p-value ($< 2e-16$), meaning we are highly confident that the intercept is different from zero.

The slope for distance is -7.583 which indicates that for every additional mile far away from the city centre, the hotel price decreases by approximately \$7.6. This suggests an **inverse relationship** between price and distance because normally the hotel gets further from the city centre, the price tends to drop. The p-value associated with this slope is low ($p < 0.001$), which is highly significant. This means that the relationship between distance and price is **statistically significant**, and we can conclude that distance has a **negative effect** on price.

The R-squared value of 0.07972 indicates that about 7.97% of the variability in hotel prices is explained by the distance from the city centre. While this shows there is a relationship, it also suggests that **other factors**, not included in this model, explain most of the variation in prices. Moreover, the adjusted R-squared also has a low value of 0.07801 similarly indicates a low explanatory variable in this model. Therefore, it suggests that the distance variable alone is not the most important factor influencing the price.

The scatter plot with a regression line:



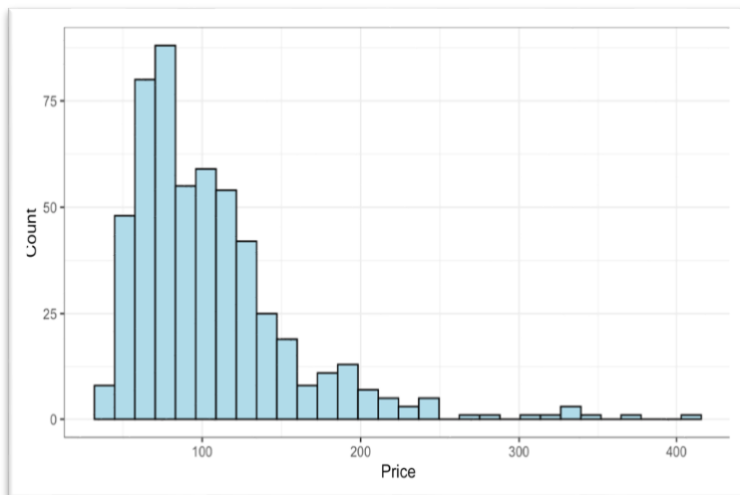
The red line (linear regression) shows a slightly negative slope, meaning that as the distance from the city centre increases, the hotel price tends to decrease. Compare the red straight line with the LOWESS curve before, we can observe the general downward trend. In comparison, the LOWESS curve shows more nuances/behaviours, especially in 0 – 2 miles range, where prices have more variability. This indicates that a simple regression line cannot fully capture the relationship as it misses some of the non-linear behaviours of the dataset. Overall, by this comparison, we can say that the relationship is not perfectly linear, a more flexible approach as LOWESS better captures the variability of the data.

5. Log transformations

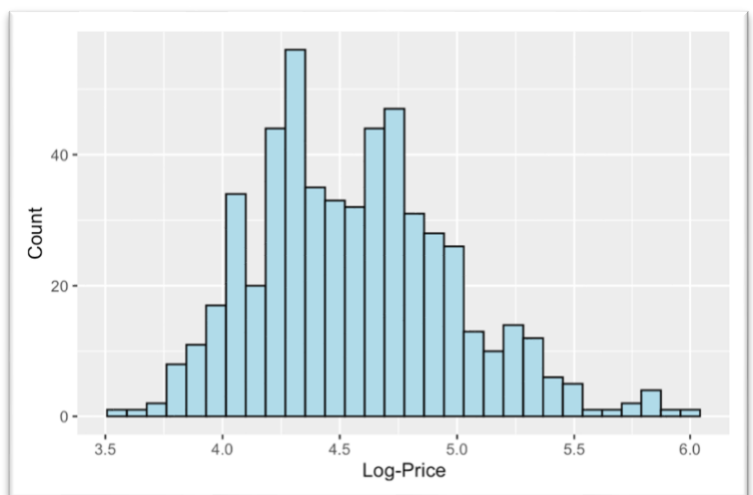
a/ Creating Log-Transformed Variables

```
# Create logged variables
city_data$ln_price <- log(city_data$price)
city_data$ln_distance <- log(city_data$distance)
```

Histograms for the **price** (before transformation) and **log-price** (after transformation)

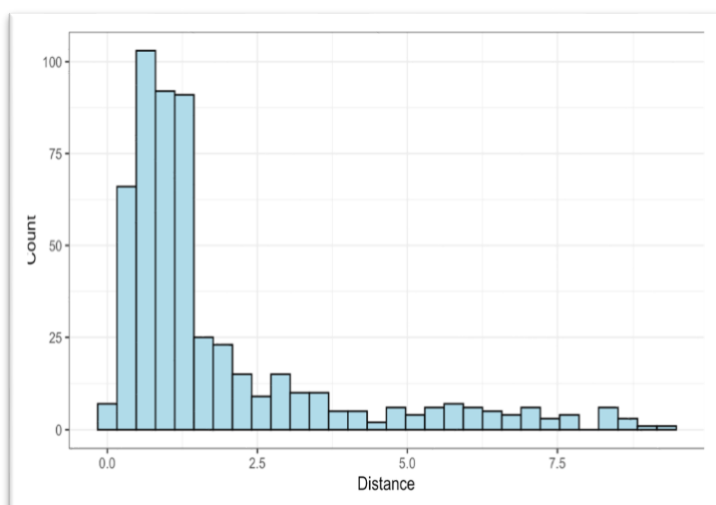


Before

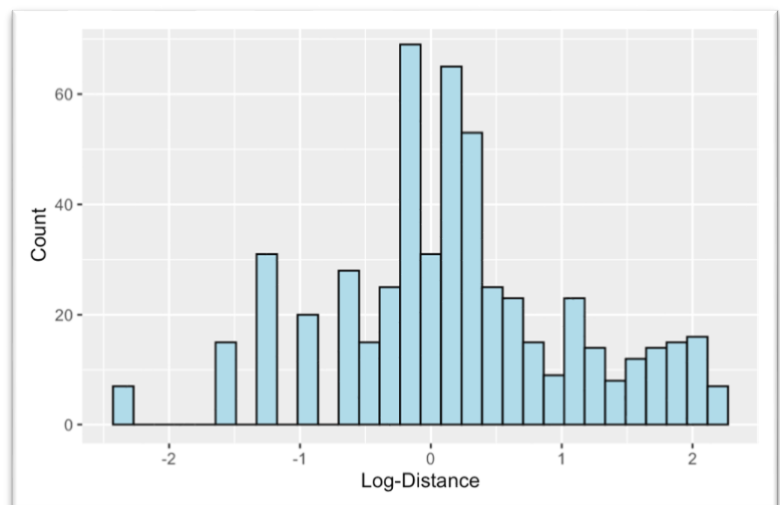


After

Histograms for the **distance** (before transformation) and **log-distance** (after transformation)



Before



After

Comments:

The histograms of the original price and distance variables shows a skewed distribution, with most values clustered at lower ranges. However, after applying the natural logarithm to both variables (price and distance), the distributions appear to be more normally distributed, reducing the skewness and bringing the data closer to the symmetrical shape.

For **log-price**: this transformed variable has a more evenly distribution compared to the original price histogram, which was heavily **skewed to the right**. The log transformation helps **normalise** the data, making it more symmetric and reducing the impact of extreme high values.

For **log-distance**: this transformed variable is more symmetric after applying logarithm in the distance variable. It is observed that it has a peak at 0, indicating that many hotels are located every near the city centre. Similarly, this transformation also reduces the skewness (right skewed in the original histogram), improving the normality of the distributions, particularly for price variable.

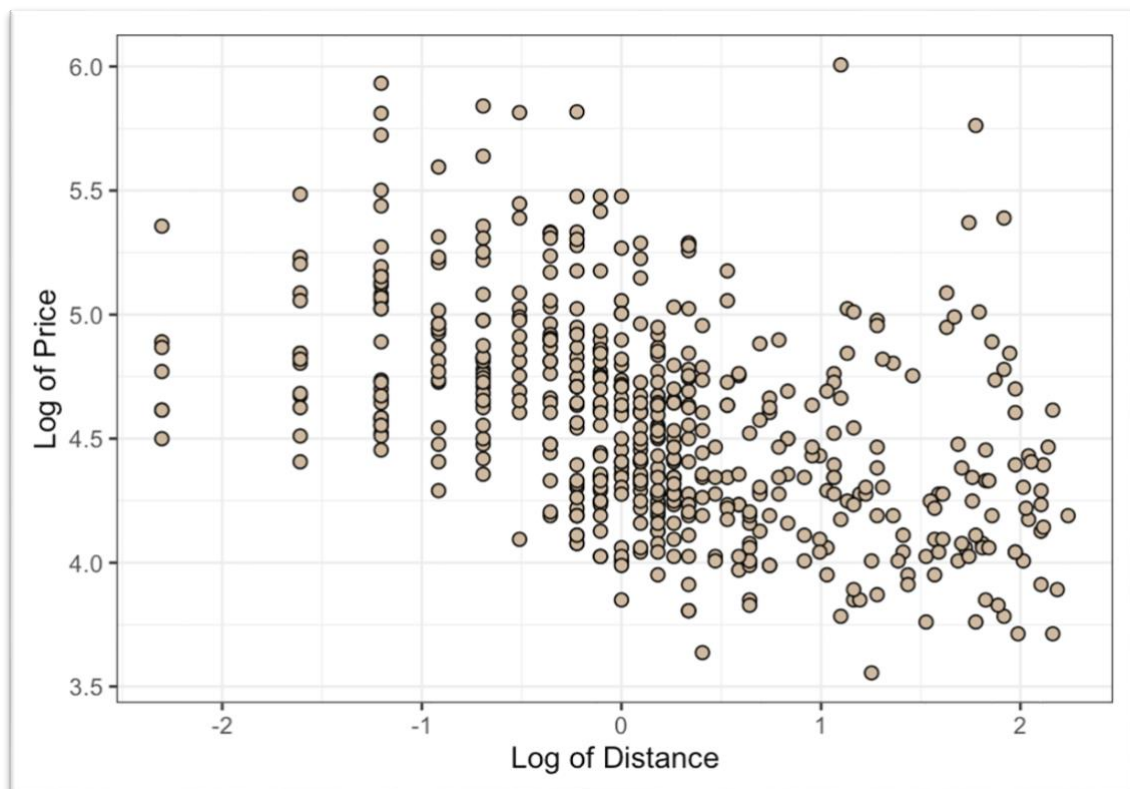
Overall, the log-transformed variables make the data more evenly distributed, and suitable for linear modelling, as they reduce the influence the outliers in price (luxury hotel prices) and distance (far away from the city centre). This is especially **useful** when performing regression analysis to make sure that the assumptions of normality are met.

b/ Scatterplots and model choice

Log-log scatter

```
# LOG_LOG scatter

ggplot(city_data, aes(x = ln_distance, y = ln_price)) +
  geom_point(fill = 'bisque3', color = 'black', shape = 21, size = 2) +
  labs(x = 'Log of Distance', y = 'Log of Price') +
  theme_bw()
```

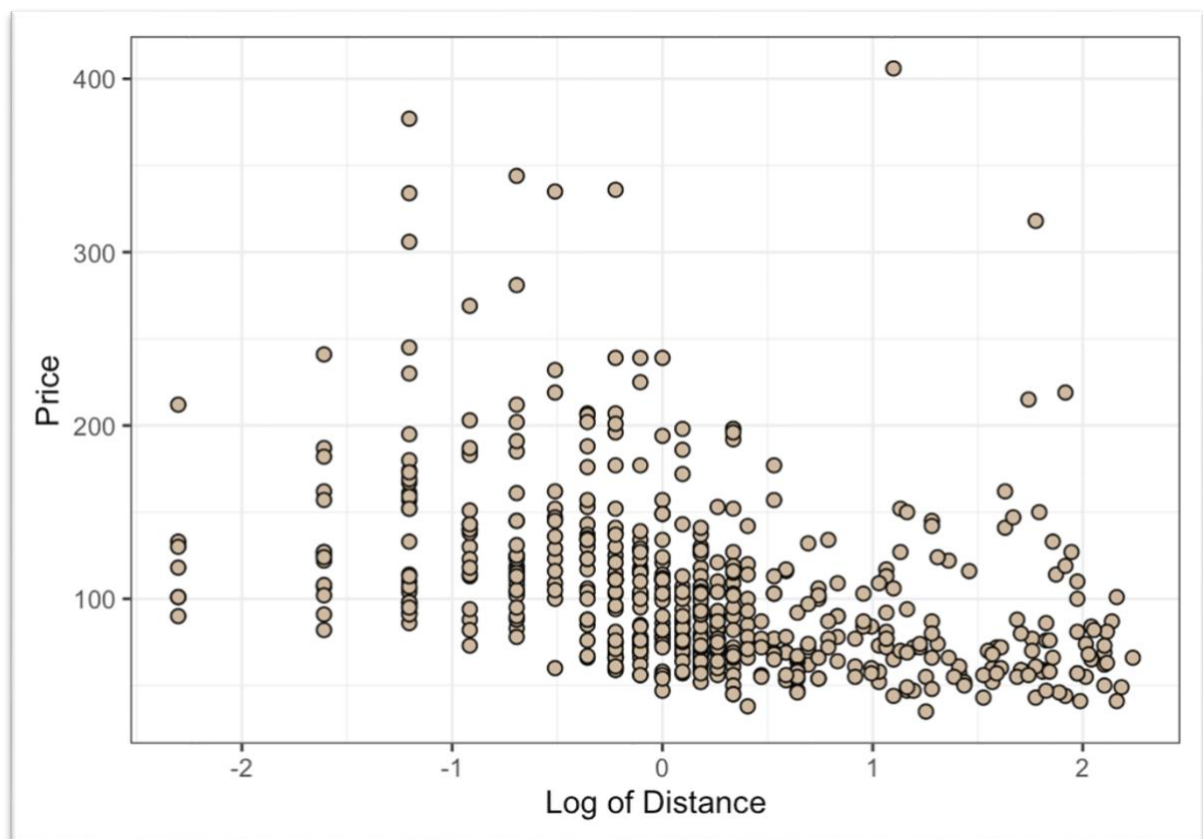


$$\text{Model: } \ln(\text{price}) = \beta_0 + \beta_1 * \ln(\text{distance}) + \epsilon$$

This log-log scatterplot provides a clear inverse/negative relationship between the distance and price. As the log distance increases, the log of price decreases, which means that hotels further from the city centre tend to have lower prices. In other words, we can say that with the **percentage changes** in distance have a proportional effect on percentage changes in price. Moreover, the pattern is fairly linear, indicating that this transformation can help in normalising the data and avoiding the skewness. It helps to smooth the data, reducing the influence of the outliers and presenting the data more uniformly.

Level-log scatter

```
ggplot(city_data, aes(x = ln_distance, y = price)) +  
  geom_point(fill = 'bisque3', color = 'black', shape = 21, size = 2) +  
  labs(x = 'Log of Distance', y = 'Price') +  
  theme_bw()
```



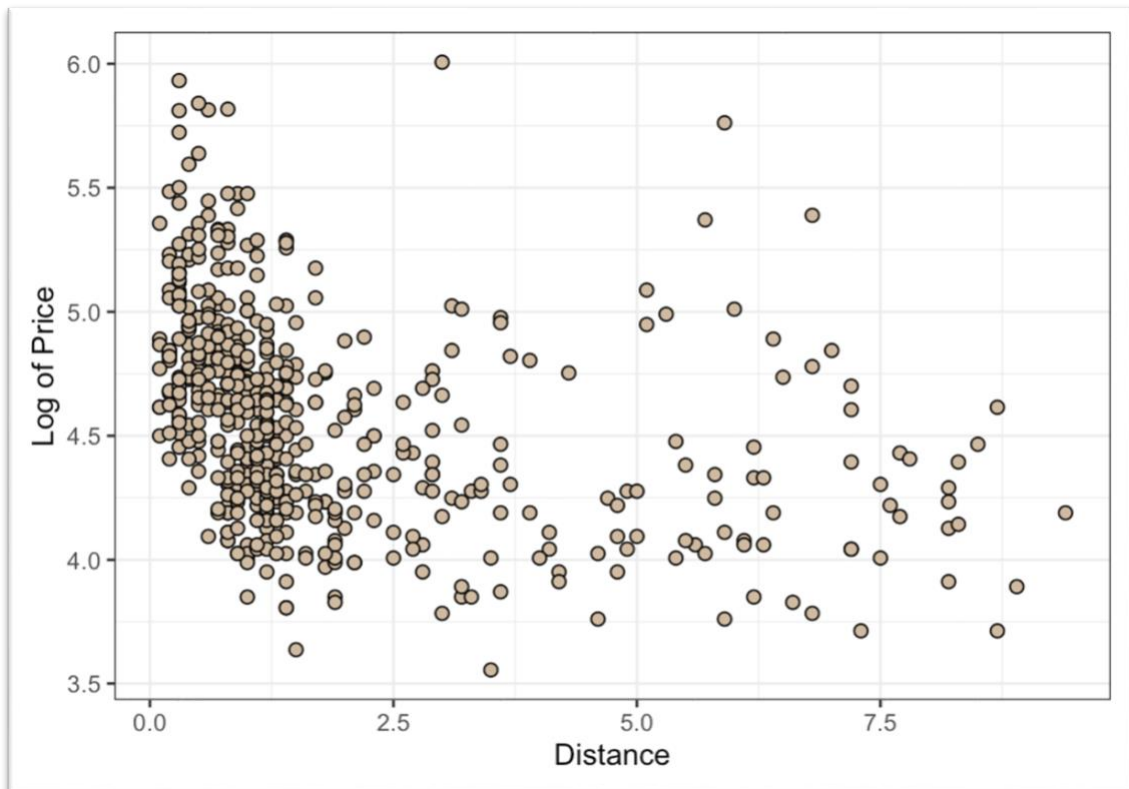
$$\text{Model: } \text{price} = \beta_0 + \beta_1 * \ln(\text{distance}) + \epsilon$$

In the level log scatterplot, the y-axis represents the actual prices (in dollars), while the x-axis is the logarithm of distance. Unlike the log-log model, this specification shows the absolute change in price as the logarithm distance increase. With this pattern displayed above, we see that as the distance from the city centre increases (on the log scale), prices generally decline. However, the relationship appears more scatter and less linear than the log-log model. Specifically, there are some high-priced outliers, particularly for the distances between 0 to 2 (on the log scale) on the x-axis, indicating a general downward trend. These outliers reduce the clarity of the relationship between the price and distance (the one we are looking at).

Overall, this model shows some limitations when it comes to the finding the relationship between the price and log of distance, especially when we must consider percentage changes in distance while the distance is not an appropriate variable to be interpreted in percent.

Log-level scatter

```
# LOG-LEVEL scatter
ggplot(city_data, aes(x = distance, y = ln_price)) +
  geom_point(fill = 'bisque3', color = 'black', shape = 21, size = 2) +
  labs(x = 'Distance', y = 'Log of Price') +
  theme_bw()
```



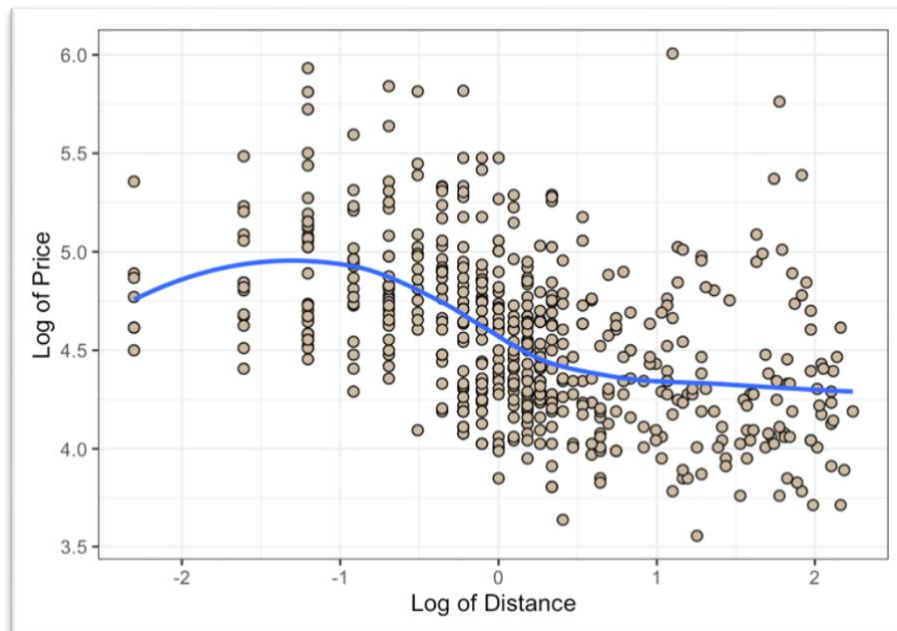
$$\text{Model: } \ln(\text{price}) = \beta_0 + \beta_1 * \text{distance} + \epsilon$$

In the log-level scatterplot, the y-axis represents the logarithms of the price, and the x-axis is the actual distance from the city centre (miles). This specification provides insights into the percentage change in price relative to the absolute distance. This pattern indicates that with the increases in the distance, there is a decrease in log(price). However, the spread is wider when it compared with the log-log model, especially at the 0-2 miles (shorter distances in the data range). The relationship between the log(price) and the distance variable is not consistent with the variation of the data points throughout the distance range, from 2.5 miles to 10 miles. Especially, there are some outliers pushing up the price at higher distances (larger than 2.5 miles).

c/ Fit a linear regression model

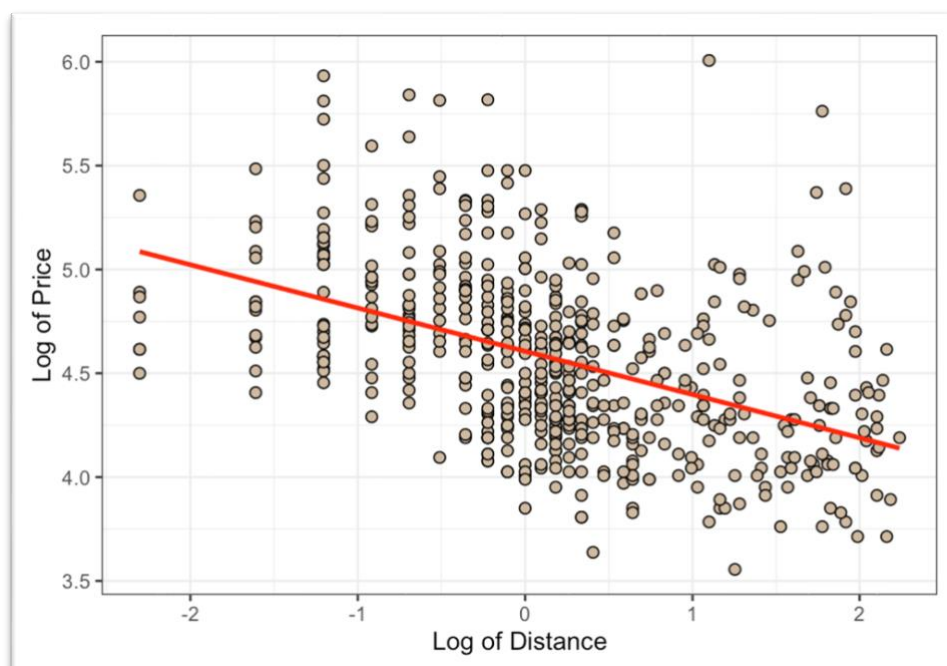
In this analysis, I choose the **log-log model** as it provides a more linear and interpretable relationship, where the percentage changes in distance correspond to percentage changes in price. This makes the model ideal for understanding how variations in distance impact price in proportional terms, especially since the price-distance relationship is nonlinear in nature. Additionally, the log-log scatterplot shows a clearer inverse relationship, whereas the distance from the city centre increases, the price tends to decrease at a proportional rate, which is consistent with economic expectations.

I am evaluating the **Loess** regression model:



The **Lowess** regression line (blue curve) provides a smoothed fit to the data. It captures the nonlinearities present in the data, showing an initial increase followed by a decline as distance increases. This non-parametric approach is useful for understanding the broader trends without assuming a strict linear relationship for these variables.

I am evaluating the **Lm** regression model:



The **linear** regression line (red line) imposes a linear fit on the log-log model. This helps to quantify the proportional relationship but may make it too simple the actual behaviours of the curve in the data. Nonetheless, it provides an indication of the general trend between distance and price.

Regression 2 table format:

Variable	Coefficient	Std. Error	t-value	p-value
Intercept	4.60596	0.01656	278.2	<2e-16
ln_distance	0.2084	0.01742	-11.96	<2e-16
	R-squared	0.21		
	Adj R squared	0.2086		

The intercept of 4.60596 indicates the predicted log of hotel prices when the log of distance is zero, which represents the base log-price for hotels located near the city centre. This intercept can be interpreted as the expected log of price for hotels in the closest proximity to the city centre. The coefficient for the log of distance (0.2084) suggests that a 1% increase in distance from the city centre leads to a 0.2084% increase in hotel prices. This positive relationship indicates that, in contrast with the typical assumptions, hotels further from the centre tend to be more expensive.

The p-values for both the intercept and the log of distance are extremely low (<2e-16), showing that the coefficients are highly statistically significant ($p < 0.001$). This means that the relationship between distance and hotel prices is unlikely due to random chance. The R-squared value of 0.21 indicates that 21% of the variation in hotel prices can be explained by the distance from the city centre. Although the R-squared value is modest, it suggests that other factors not included in the model are influencing hotel pricing, meaning that distance alone does not capture all the variability in price.

6. Linear piecewise spline

Fitting the linear piecewise spline model in levels no logs with 1 knot (cutoff point).

The linear piecewise spline model for the level-level (x = distance, y = price) model, with the cutoff = 2 because after 2 miles distance, the relationship between price and distance less steep and kind of flat which is different from the first range of distance (drop significantly). This distance would be my knot (at 2 miles).

Regression 3 table format: (for the level-level mode)

Variable	Coefficient	Std. Error	t-value	p-value
Intercept	151.037	4.854	31.118	<2e-16
lspline(distance, cutoff)1	-38.848	4.115	-9.441	<2e-16
lspline(distance, cutoff)2	2.208	1.631	1.354	1.76E-01
R-squared	0.1747			
Adj R squared	0.1716			

In the spline regression model (Regression 3), the intercept is 151.037, representing the predicted hotel price (\$151.037) at the starting point of the distance variable (close to the city centre). The coefficient for the first segment of the distance spline (-38.848) shows that hotel prices drop by approximately \$38.85 for each unit increase in distance from the city centre within this range. This coefficient is statistically significant, as indicated by the extremely low p-value (<2e-16).

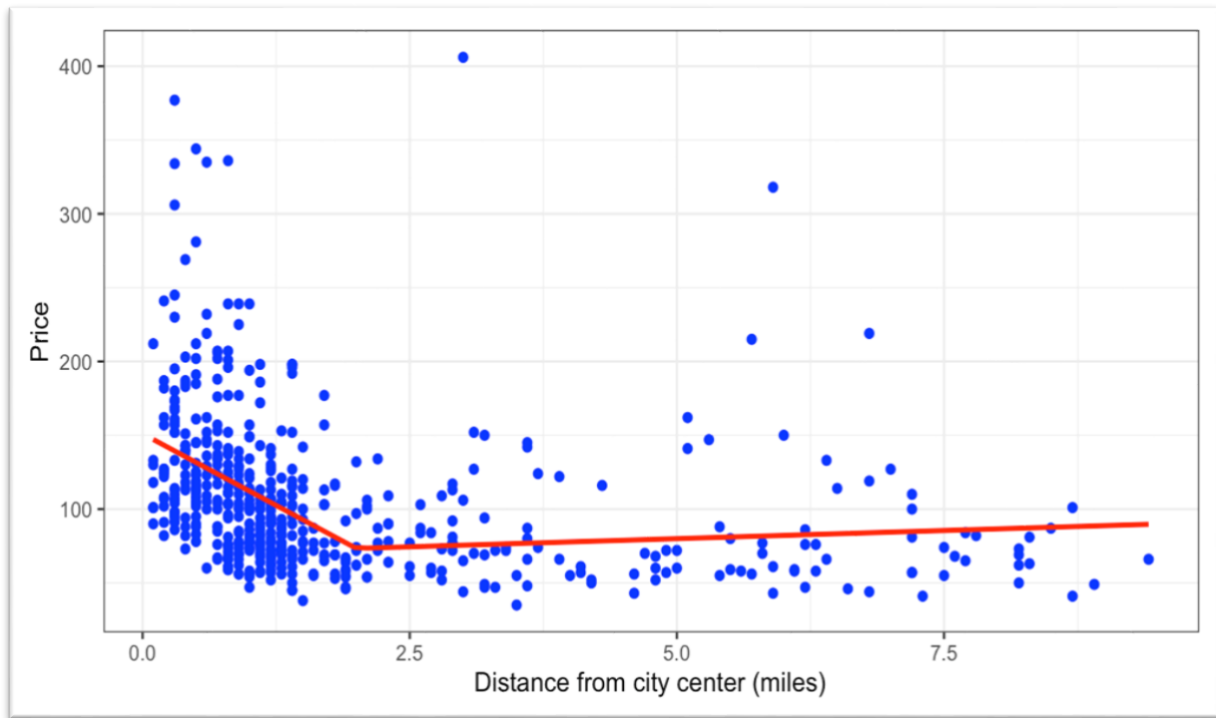
The second spline segment coefficient (2.208) suggests that, beyond the cutoff point, hotel prices increase by \$2.21 for every additional mile from the city centre. However, the p-value (0.176) for this coefficient indicates that this positive relationship is not statistically significant at common significance levels. This could suggest that hotel prices have not a strong relationship to distance.

The R-squared value of 0.1747 implies that 17.47% of the variability in hotel prices is explained by the distance from the city centre, with the spline approach providing a more nuanced fit compared to a simple linear model. However, other factors are likely influencing the price, given the relatively low R-squared value. The spline model provides flexibility in capturing different trends across distance ranges.

I consider the dots' colour for better visualisation:

```
# Create scatterplot with regression line

ggplot(data = city_data, aes(x = distance, y = price)) +
  geom_point(color = "blue") +
  geom_smooth(formula = y ~ lspline(x, cutoff), method = "lm", color = "red", se = FALSE) +
  labs(x = "Distance from city center (miles)", y = "Price") +
  theme_bw()
```



The scatter plot illustrates the relationship between hotel prices and distance from the city centre, with a clear downward trend in prices for hotels within 2.5 miles of the centre. The spline regression line (red) reflects this, showing a sharp decline initially (with a coefficient of -38.848) as distance increases. Beyond 2.5 miles, the line flattens and slightly rises (coefficient of 2.208), indicating that prices stabilize or increase slightly at greater distances. The spline model effectively captures this non-linear pattern across different distance ranges.

I argue with 1 more model with linear piecewise spline **Log-level** model ($x = \text{distance}$, $y = \ln_{\text{price}}$) to see how the data works in this model, and I call it as reg4.

Regression 4 table format (for the log-level model)

Variable	Coefficient	Std. Error	t-value	p-value
Intercept	4.982374	0.037899	131.466	<2e-16
lspline(distance, cutoff)1	-0.351355	0.03213	-10.935	<2e-16
lspline(distance, cutoff)2	0.008342	0.012736	0.655	0.0.513
R-squared	0.2367			
Adj R squared	0.2338			

In Regression 4, the intercept is 4.982374, indicating that when the distance from the city centre is 0, the log of the hotel price is approximately 4.98 (which is defined to a price around $e^{4.982374} = \$146$). The first spline coefficient (-0.351355) suggests that for distances up to the cutoff (at 2 miles), every additional mile from the city centre decreases the log price by about 0.35, corresponding to a decrease in actual hotel prices. This is highly significant with a p-value less than $2e-16$, indicating a strong relationship for this portion of the data.

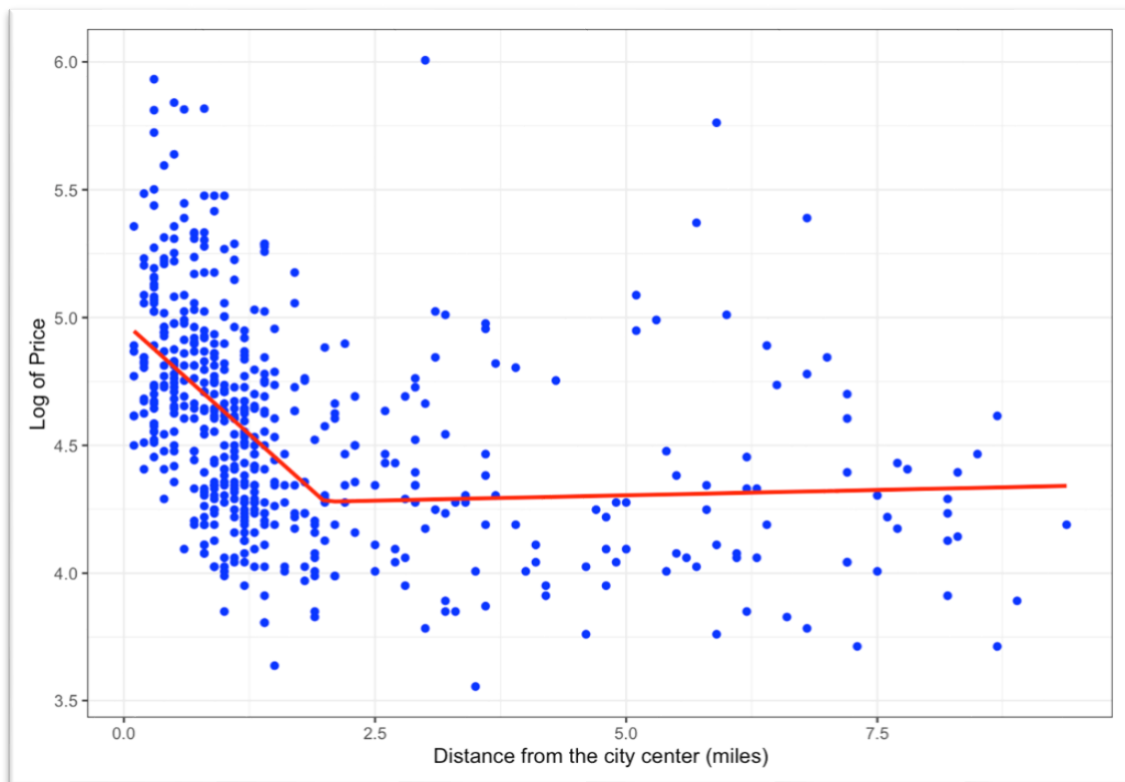
However, the second spline coefficient (0.008342) is not significant, as indicated by its high p-value of 0.513. This suggests that beyond the cutoff point, the change in log prices with distance is not statistically significant, indicating that prices stabilise after a certain distance. The R-squared value of 0.2367 shows that around 23.67% of the variation in hotel prices is explained by the distance from the city centre, which is an improvement compared to previous models, but might still need more variables involved for further investigation.


```
# !!! I argue with 1 more model: log-level

# Estimate the model and call it reg4
reg4 <- lm(ln_price ~ lspline(distance, cutoff), data = city_data)

summary(reg4) # display the regression output

ggplot(data=city_data, aes(x = distance, y = ln_price)) +
  geom_point(color = "blue") +
  geom_smooth(formula = y ~ lspline(x, cutoff), method = "lm", color = "red", se = FALSE) +
  labs(x = "Distance from the city center (miles)", y = "Log of Price") +
  theme_bw()
```



The scatterplot visualises this spline regression, with the sharp decrease in prices for distances up to the cutoff (cutoff point = 2), and then a stabilisation beyond that point. The red line represents the spline regression's segmented effect, clearly showing the behaviour of the data point before and after the cutoff point.

I would choose the piecewise spline for the log-level model (reg4) for further investigation because it has the highest R-squared value (0.2367) compared to the normal spline regression level-level model (0.21). This suggests that the spline for log-level model explains more of the variation in the dataset, providing a better fit. Specifically, a higher R-squared value indicates that the model better captures the relationship between distance and hotel price, leading to a more accurate interpretation of the fitted line.

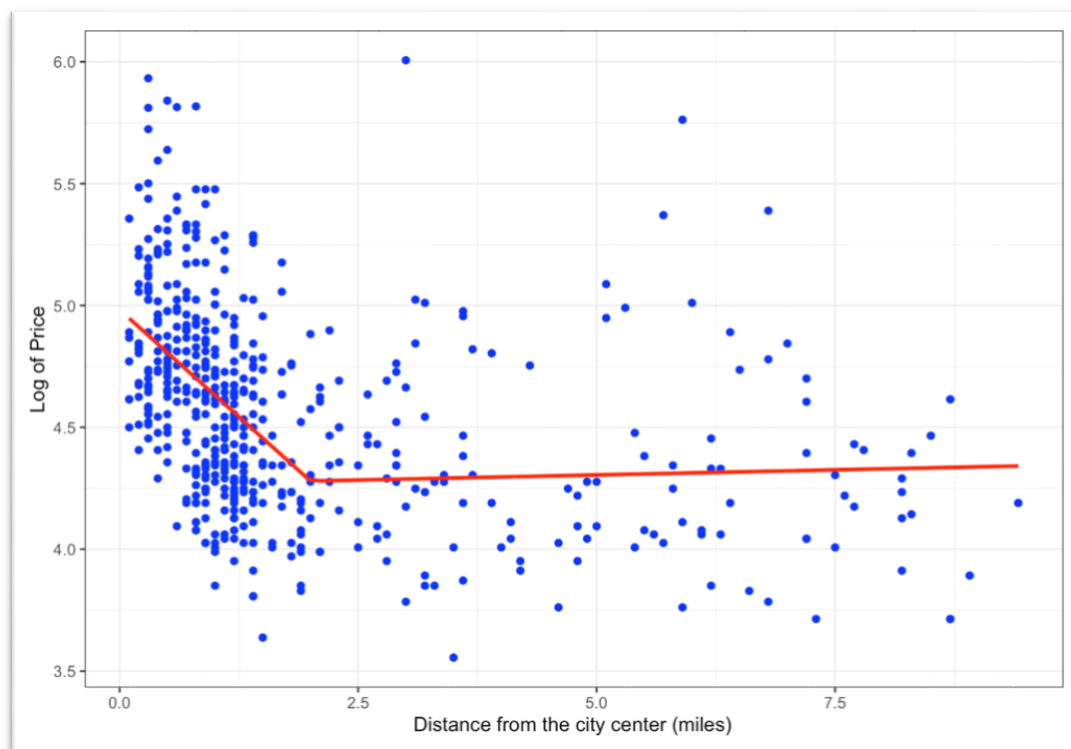
7. Choosing the best model

Choosing the best model between a linear regression model, log-log model (transformations), and piecewise linear spline model as reg1, reg2 and reg4 respectively using `msummary(list(reg1, reg2, reg4))` command and the result for each regression is displayed below in respectively:

	(1)	(2)	(3)
(Intercept)	120.063	4.606	4.982
	(2.989)	(0.017)	(0.038)
distance	-7.583		
	(1.111)		
ln_distance		-0.208	
		(0.017)	
lspline(distance, cutoff)1			-0.351
			(0.032)
lspline(distance, cutoff)2			0.008
			(0.013)
Num.Obs.	540	540	540
R2	0.080	0.210	0.237
R2 Adj.	0.078	0.209	0.234
AIC	5764.3	483.3	466.8
BIC	5777.2	496.2	483.9
Log.Lik.	-2879.174	-238.646	-229.387
F	46.603	143.037	83.244
RMSE	50.04	0.38	0.37

The piecewise spline model in regression 4 (model 3) is the best choice for the prediction of the relationship between distance and hotel prices, as it has the highest R-squared (0.237), indicating it explains the most variability in the data compared to the linear (0.078) and log-log (0.210) models. The spline model effectively captures nonlinear patterns (the more nuances/ behaviours of the data) by allowing for different relationships at specific cutoff points (at 2 miles), making it more flexible in modelling the real-world trends between distance and price. Additionally, with its lower AIC, BIC, and RMSE values, it provides a more efficient and accurate prediction, highlighting its superiority over simpler models. This combination of flexibility, accuracy, and interpretability makes regression 4 (model 3) the best model in this analysis.

I choose the spline for the log-level model as below:



Command to compute residuals and fitted values:

```
city_data$lnprice_hat <- reg4$fitted.values # Fitted (predicted) values
city_data$lnprice_resid <- reg4$residuals   # Residuals
```

The fitted values (lnprice_hat) represent the predicted log prices generated by the regression model, while the residuals (lnprice_resid) capture the difference between the **actual** log prices and the **predicted** values. These residuals help assess how well the data points in the model fit the data, where **smaller residuals indicate a better fit**. Analysing the distribution of residuals can also help detect any systematic patterns or model deficiencies, such as nonlinearity or heteroscedasticity, that may require further model adjustments.

Run the command below to get the hotels that are the most **UNDERPRICED** (top deals = **green**):

```
# Get the hotels that are the most UNDERPRICED
city_data |>
  slice_min(lnprice_resid, n = 5) |> # Select the row with the n smallest residuals
  select(hotel_id, price, distance, lnprice_hat, lnprice_resid) # Only interested in specific columns
```

In this step, I used the *slice_min* function to identify the top 5 underpriced hotels by selecting the rows with the **smallest residuals** (in other words, it can be called as largest negative). The smallest residuals indicate that the actual price is much lower than the predicted price, meaning that it is a better deal or underpriced relative to what the model would expect. Essentially, a small residual (close to 0 or negative) suggests that the hotel is priced lower than what the model predicts, making it a **good deal** for customers looking for lower prices. The residuals (lnprice_resid) represent the difference between the actual log price and the predicted log price. A smaller residual indicates that the hotel is underpriced compared to the model's prediction. The result highlights the most underpriced hotels, marked as 'top deals' in green in the visualisation.

BEST 5 (UNDERPRICED): Actual Prices < Predicted Prices

	hotel_id	price	distance	lnprice_hat	lnprice_resid
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	<u>19640</u>	38	1.5	4.46	-0.818
2	<u>17526</u>	47	1	4.63	-0.781
3	<u>15186</u>	35	3.5	4.29	-0.737
4	<u>17762</u>	45	1.4	4.49	-0.684
5	<u>17761</u>	45	1.4	4.49	-0.684

In the result table, it shows that there are a few hotels (19640, 17526, 15186, 17762 and 17761) have the lower prices than predicted (with the negative residuals). The prices of these hotels are in the range of less than \$100 with the distance of around 1.5 miles far from the city centre. Noticabale, there are 2 hotels (different hotel ids) have the same price (\$45) so it could create an overlap data point when I run the command for the scatterplot. Overall, those hotels could be in use for the further recommendations for tourists who are likely to stay near the city centre with the reasonable budget.

Run the command below to get the hotels that are the most **OVERPRICED**:

```
# Get the hotels that are the most OVERPRICED
city_data |>
  slice_max(lnprice_resid, n = 5) |> # Select the row with the n largest residuals
  select(hotel_id, price, distance, lnprice_hat, lnprice_resid) # Only interested in specific columns
```

WORST 5 (OVERPRICED): Actual Prices > Predicted Price

	hotel_id	price	distance	lnprice_hat	lnprice_resid
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	<u>18257</u>	406	3	4.29	1.72
2	<u>15351</u>	318	5.9	4.31	1.45
3	<u>15623</u>	336	0.8	4.70	1.12
4	<u>15353</u>	219	6.8	4.32	1.07
5	<u>15793</u>	215	5.7	4.31	1.06

For the overpriced hotels (with the largest residuals or positive values), we observe that the actual prices exceed the predicted prices significantly. These hotels have large residual values, indicating that they are priced much higher than the model estimates based on their distance and other factors. This suggests that these hotels may not provide good value for money, as their prices are not justified by the model's predictions, making them less attractive deals for potential customers. The slice_max function is used to extract the rows with the largest residuals, highlighting these overvalued hotels (18257, 15351, 15623, 15353, and 15793). Specifically, those hotel prices are range from over \$200 to around \$400 and the hotel id 198257 has the highest price at \$406.

Command to create the **top5** and **bottom5** indicator variables using *ifelse()* function

```
# Create bottom5 and top5 indicator variables using ifelse()
city_data$deal_indicator <- ifelse(city_data$lnprice_resid %in% head(sort(city_data$lnprice_resid, decreasing = FALSE), 5), 'top5',
  ifelse(city_data$lnprice_resid %in% head(sort(city_data$lnprice_resid, decreasing = TRUE), 5), 'bottom5', 'other'))
```

The *ifelse()* function in this command is used to create a new variable called *deal_indicator*, which categorises the hotels into three groups: "**top5**" in **green**, "**bottom5**" in **red**, and "**other**" as required color is **gray**. The top5 category includes the hotels with the smallest residuals (best deals), while the bottom5 category includes the hotels with the largest residuals (worst deals) and those prices are similarly with the same idea. All other hotels are categorised as "**other**". This classification helps identify the most underpriced and overpriced hotels based on their residual values from the regression model. In real life, visualising the best and worst deals, as shown in this scatter plot, provides valuable insights into pricing anomalies. We can deliver the best insights into the data points as we have to predict, recommend and adjust for further investigation.

Scatter plot with top5 best (**UNDERPRICED**) deals in **green** and bottom5 worst (**OVERPRICED**) deals in **red**.



This scatter plot presents an insightful analysis of hotel pricing in relation to distance from the city centre, highlighting both underpriced and overpriced deals. The top5 underpriced deals (shown in **green**) reflect hotels where the actual price is significantly lower than the predicted price based on their distance, offering customers excellent value for money. These hotels are located closer to the city centre (within approximately 2 miles), which makes the lower price even more attractive considering their prime location.

On the other hand, the bottom 5 overpriced deals (in **red**) highlight hotels where the actual price greatly exceeds the predicted value. These hotels are mostly situated further from the city centre (beyond 2.5 miles), and yet their prices are high. This suggests that they are poor deals for customers, as the price charged does not align with the market expectations for their distance from the city.

The visual contrast between the green and red points allows us to clearly see which hotels varies most from their predicted values, then guiding decision-makers in identifying pricing anomalies. In real-world applications, such insights could be used by hotel managers or pricing strategists to adjust rates for better market alignment and to ensure they offer competitive, value-for-money deals to attract customers.

CONCLUSION

Overall, the analysis of hotel prices in relation to distance from the city centre reveals a clear **inverse relationship**, where hotels closer to the city centre tend to have higher prices (in contrast with their relationship in real life). This effect is most observed within the first 2 miles from the centre, after which prices stabilise. The piecewise spline model (log-level) with a cutoff point at 2 miles proved to be the most suitable model, with an R-squared of 0.237, explaining a significant portion of the variation in hotel prices based on distance variable. The model effectively captures the **non-linear** pattern observed in the data, where the impact of distance diminishes beyond a certain point.

However, the model has its **limitations**. It primarily relies on distance as the key predictor of hotel prices and does not consider other factors that may influence pricing, such as hotel star ratings, customer reviews, or demand fluctuations due to low and peak travel season. As a result, the model may too be simple to deal with the complex pricing mechanisms in the hospitality industry. Additionally, the R-squared value, while the highest among the models tested, still suggests that a significant amount of variability in hotel prices remains unexplained. Moreover, the R-squared value measures how well the independent variable(s) explain the variation in the dependent variable. It tells us the proportion of the variance in the dependent variable that is predictable from the independent variables. It helps indicate the model's goodness-of-fit, showing how much of the outcome variability is explained by the model only. When it comes to the relationship between these 2 variables, other tools could be used for better capture at most variations of the data, not only relying on the model with only distance and R-squared.

To improve the model, I recommend including additional variables that better capture hotel quality and customer preferences. Variables like star ratings, available amenities, and user reviews could add more explanatory power to the model. Furthermore, using more advanced models, such as generalised additive models (GAMs), could help capture non-linear relationships between multiple predictors and hotel prices. Interaction terms, such as the interactions between distance and hotel quality, could also provide deeper insights into how different factors jointly affect pricing. These adjustments might yield a more accurate and reliable model for understanding hotel pricing dynamics.