Q3

Marcus Martinez

2024-08-17

Q3

Set Up

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
data <- read_csv("/Users/navyasinghal/Desktop/Summer/Intro to Machine Learning/Unsupervised Learning/ST
```

```
## Rows: 7894 Columns: 23

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (23): CS_PropertyID, cluster, size, empl_gr, Rent, leasing_rate, stories...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(data)
```

```
## spc_tbl_ [7,894 x 23] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ CS_PropertyID    : num [1:7894] 379105 122151 379839 94614 379285 ...
##  $ cluster          : num [1:7894] 1 1 1 1 1 1 6 6 6 6 ...
##  $ size             : num [1:7894] 260300 67861 164848 93372 174307 ...
```

```
##  $ empl_gr          : num [1:7894] 2.22 2.22 2.22 2.22 2.22 2.22 4.01 4.01 4.01 4.01 ...
##  $ Rent             : num [1:7894] 38.6 28.6 33.3 35 40.7 ...
##  $ leasing_rate     : num [1:7894] 91.4 87.1 88.9 97 96.6 ...
##  $ stories          : num [1:7894] 14 5 13 13 16 14 11 15 31 21 ...
##  $ age              : num [1:7894] 16 27 36 46 5 20 38 24 34 36 ...
##  $ renovated        : num [1:7894] 0 0 1 1 0 0 0 0 0 1 ...
##  $ class_a          : num [1:7894] 1 0 0 0 1 1 0 1 1 1 ...
##  $ class_b          : num [1:7894] 0 1 1 1 0 0 1 0 0 0 ...
##  $ LEED             : num [1:7894] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Energystar       : num [1:7894] 1 0 0 0 0 0 1 0 0 0 ...
##  $ green_rating     : num [1:7894] 1 0 0 0 0 0 1 0 0 0 ...
##  $ net              : num [1:7894] 0 0 0 0 0 0 0 0 0 0 ...
##  $ amenities        : num [1:7894] 1 1 1 0 1 1 1 1 1 1 ...
##  $ cd_total_07      : num [1:7894] 4988 4988 4988 4988 4988 ...
##  $ hd_total07       : num [1:7894] 58 58 58 58 58 58 1670 1670 1670 1670 ...
##  $ total_dd_07      : num [1:7894] 5046 5046 5046 5046 5046 ...
##  $ Precipitation    : num [1:7894] 42.6 42.6 42.6 42.6 42.6 ...
##  $ Gas_Costs        : num [1:7894] 0.0137 0.0137 0.0137 0.0137 0.0137 ...
##  $ Electricity_Costs: num [1:7894] 0.029 0.029 0.029 0.029 0.029 ...
##  $ cluster_rent     : num [1:7894] 36.8 36.8 36.8 36.8 36.8 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   CS_PropertyID = col_double(),
##   ..   cluster = col_double(),
##   ..   size = col_double(),
##   ..   empl_gr = col_double(),
##   ..   Rent = col_double(),
##   ..   leasing_rate = col_double(),
##   ..   stories = col_double(),
##   ..   age = col_double(),
##   ..   renovated = col_double(),
##   ..   class_a = col_double(),
##   ..   class_b = col_double(),
##   ..   LEED = col_double(),
##   ..   Energystar = col_double(),
##   ..   green_rating = col_double(),
##   ..   net = col_double(),
##   ..   amenities = col_double(),
##   ..   cd_total_07 = col_double(),
##   ..   hd_total07 = col_double(),
##   ..   total_dd_07 = col_double(),
##   ..   Precipitation = col_double(),
##   ..   Gas_Costs = col_double(),
##   ..   Electricity_Costs = col_double(),
##   ..   cluster_rent = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(data)
```

```
## # A tibble: 6 x 23
##   CS_PropertyID cluster   size empl_gr  Rent leasing_rate stories   age
##           <dbl>   <dbl>  <dbl>   <dbl> <dbl>        <dbl>   <dbl> <dbl>
## 1        379105       1 260300    2.22  38.6         91.4      14    16
```

```
## 2        122151       1  67861    2.22  28.6         87.1       5    27
## 3        379839       1 164848    2.22  33.3         88.9      13    36
## 4         94614       1  93372    2.22  35           97.0      13    46
## 5        379285       1 174307    2.22  40.7         96.6      16     5
## 6         94765       1 231633    2.22  43.2         92.7      14    20
## # i 15 more variables: renovated <dbl>, class_a <dbl>, class_b <dbl>,
## #   LEED <dbl>, Energystar <dbl>, green_rating <dbl>, net <dbl>,
## #   amenities <dbl>, cd_total_07 <dbl>, hd_total07 <dbl>, total_dd_07 <dbl>,
## #   Precipitation <dbl>, Gas_Costs <dbl>, Electricity_Costs <dbl>,
## #   cluster_rent <dbl>
```

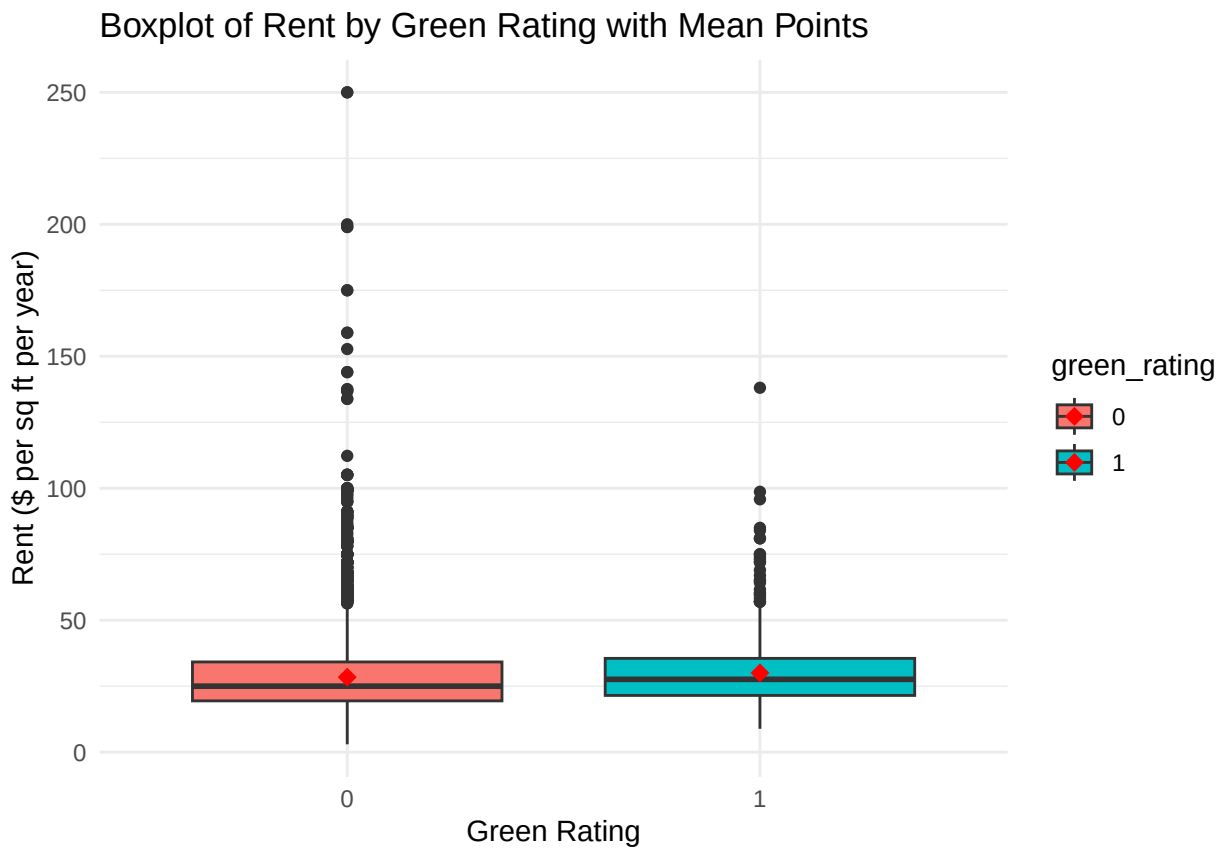
```r
# Encode categorical variables
data <- data %>%
  mutate(
         CS_PropertyID <- as.factor(CS_PropertyID),
         cluster <- as.factor(cluster),
         renovated <- as.factor(renovated),
         class_a <- as.factor(class_a),
         class_b <- as.factor(class_b),
         green_rating = as.factor(green_rating),
         LEED = as.factor(LEED),
         Energystar = as.factor(Energystar),
         net = as.factor(net),
         amenities = as.factor(amenities))
data$class = if_else(data$class_a == 1, "Class A", if_else(data$class_b == 1, "Class B", "Class C"))

# Remove buildings with very low occupancy rates (< 10%)
data <- data %>% filter(leasing_rate >= 0.1)
```

Reasons to remove buildings with very low occupancy rates (< 10%)

- **Focus on Relevant Data** - Removing buildings with very low occupancy rates before the confounding analysis ensures that we are analyzing a more relevant subset of buildings which makes the confounding analysis more meaningful.
- **Reduced Noise** - Low-occupancy buildings may introduce noise into your analysis, making it harder to discern true relationships between variables like green certification and rent.

Look at Rent

```r
ggplot(data, aes(x = green_rating, y = Rent, fill = green_rating)) +
  geom_boxplot() +
  stat_summary(fun = mean, geom = "point", color = "red", size = 3, shape = 18) +
  labs(title = "Boxplot of Rent by Green Rating with Mean Points", x = "Green Rating", y = "Rent ($ per
  theme_minimal()
```

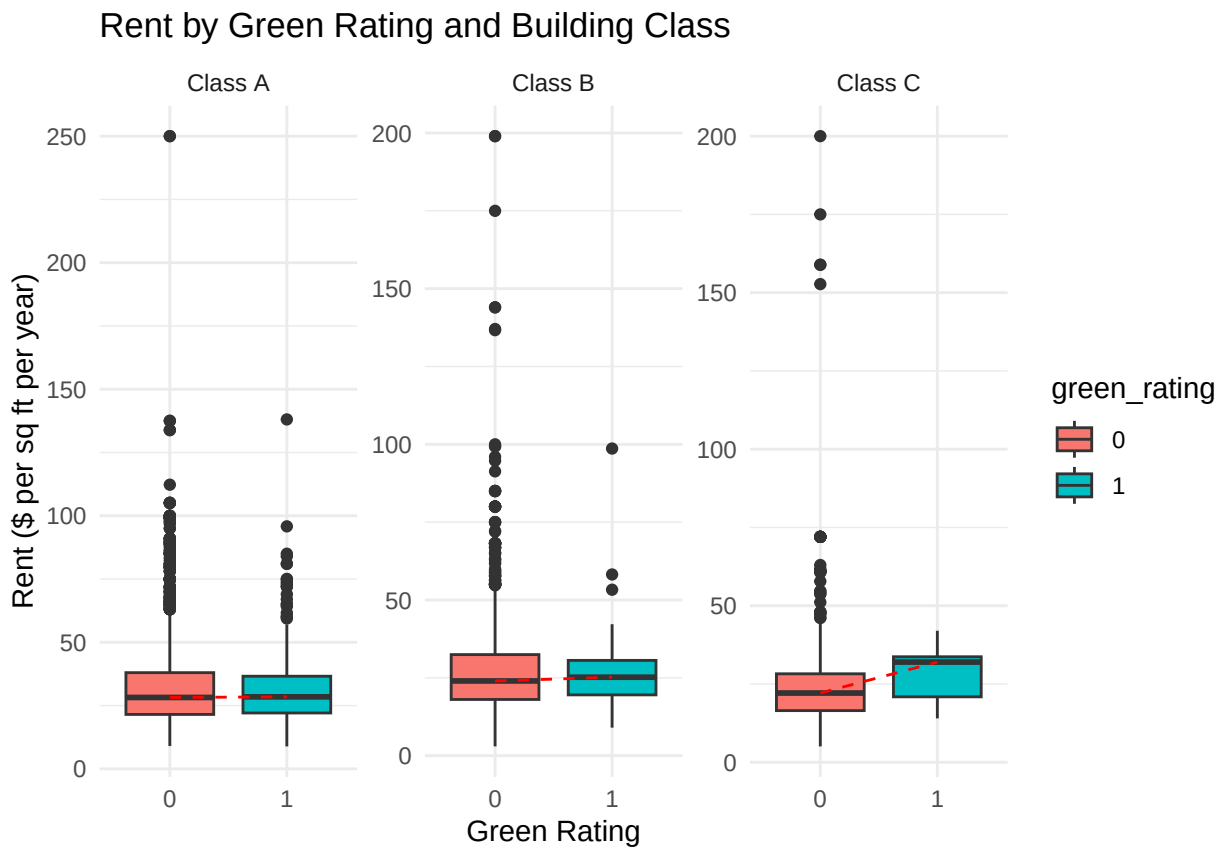


Insights - The boxplot shows that green-certified buildings generally have higher median rents compared to non-certified buildings. The addition of red mean points further highlights this trend. However, the overlap in the interquartile ranges suggests that while green certification does contribute to higher rents, other factors might also be influencing rent, making green certification just one of several considerations.

Rent by Class of Building

```
ggplot(data, aes(x = green_rating, y = Rent, fill = green_rating)) +
  geom_boxplot() +
  stat_summary(fun.y = median, geom = "line", aes(group = 1), color = "red", linetype = "dashed") +  # A
  labs(title = 'Rent by Green Rating and Building Class', x = 'Green Rating', y = 'Rent ($ per sq ft per
  facet_wrap(~ class, scales = "free") +
  theme_minimal()
```

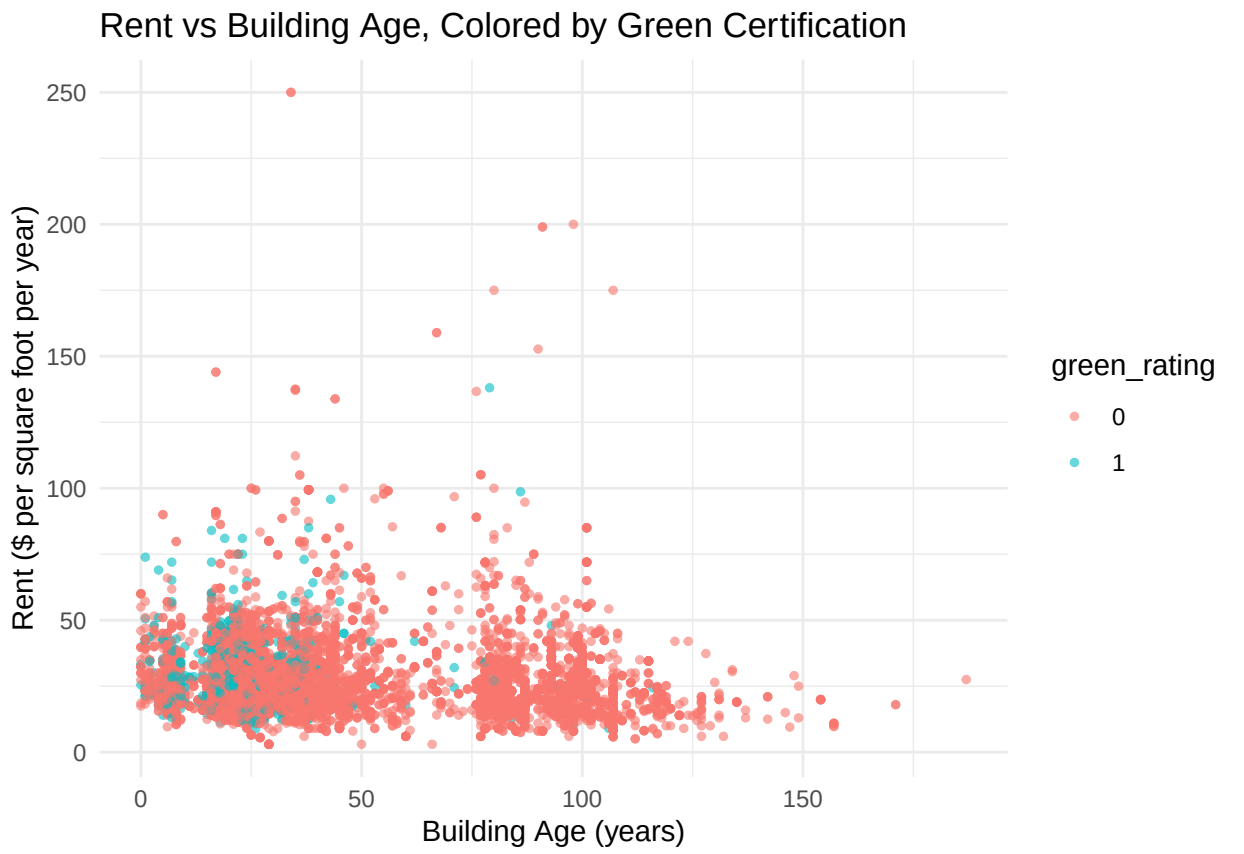
```
## Warning: The ‘fun.y‘ argument of ‘stat_summary()‘ is deprecated as of ggplot2 3.3.0.
## i Please use the ‘fun‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



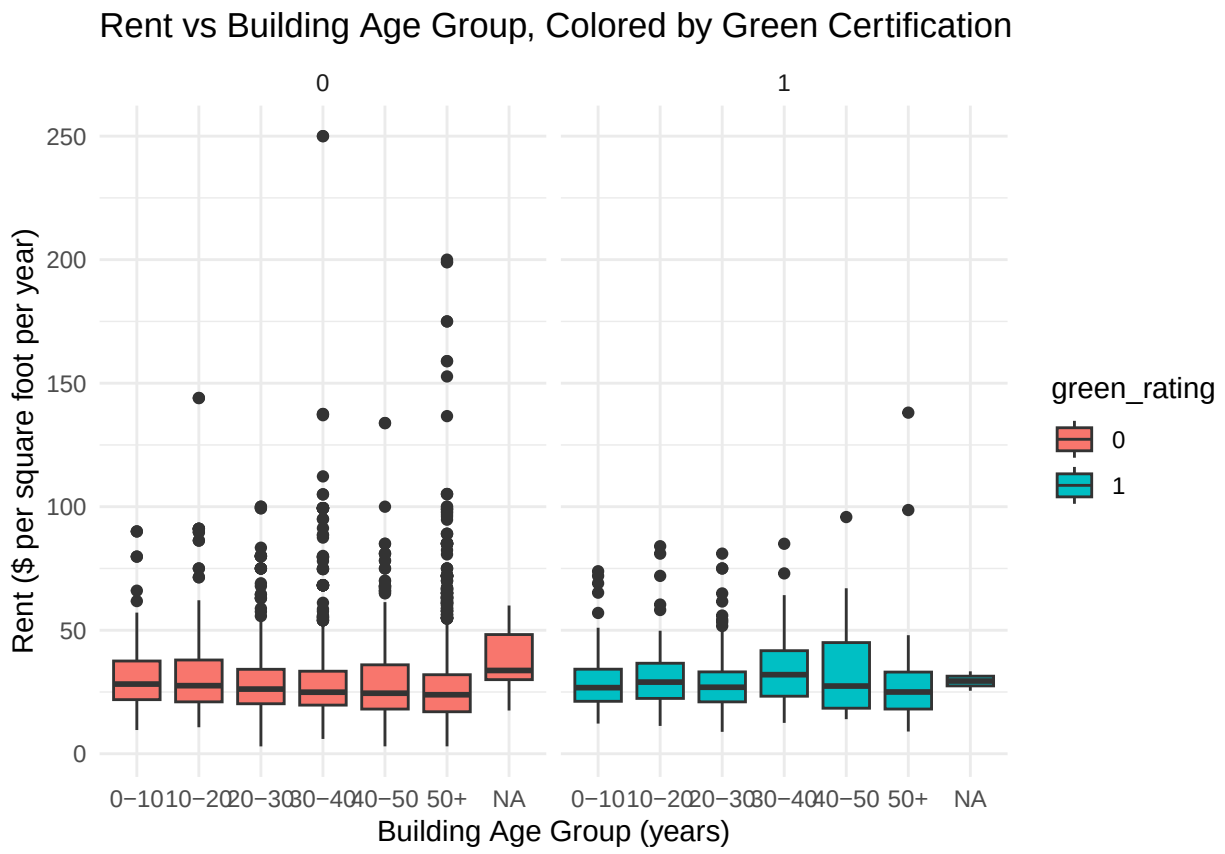
Insights - The boxplot faceted by building class (Class A, B, C) reveals that the impact of green certification on rent varies by class. In Class A buildings, the median rent difference between certified and non-certified buildings is minimal, as indicated by the red dashed median line. However, in Class B and C buildings, green certification seems to contribute more noticeably to higher rents, suggesting that green certification might be a more distinguishing factor in lower-tier buildings where tenants may value the added benefits more.

Rent by Age of the building

```
# Scatterplot: Rent vs Building Age colored by Green Certification
ggplot(data, aes(x = age, y = Rent, color = green_rating)) +
  geom_point(alpha = 0.6, size = 1) +  # Use smaller points
  labs(title = "Rent vs Building Age, Colored by Green Certification",
       x = "Building Age (years)",
       y = "Rent ($ per square foot per year)") +
  theme_minimal()
```




```
# Create age bins
data$age_group <- cut(data$age, breaks = c(0, 10, 20, 30, 40, 50, Inf),
                              labels = c("0-10", "10-20", "20-30", "30-40", "40-50", "50+"))

# Boxplot of Rent vs. Age Group, Faceted by Green Certification
ggplot(data, aes(x = age_group, y = Rent, fill = green_rating)) +
  geom_boxplot() +
  labs(title = "Rent vs Building Age Group, Colored by Green Certification",
       x = "Building Age Group (years)",
       y = "Rent ($ per square foot per year)") +
  facet_wrap(~ green_rating) +
  theme_minimal()
```



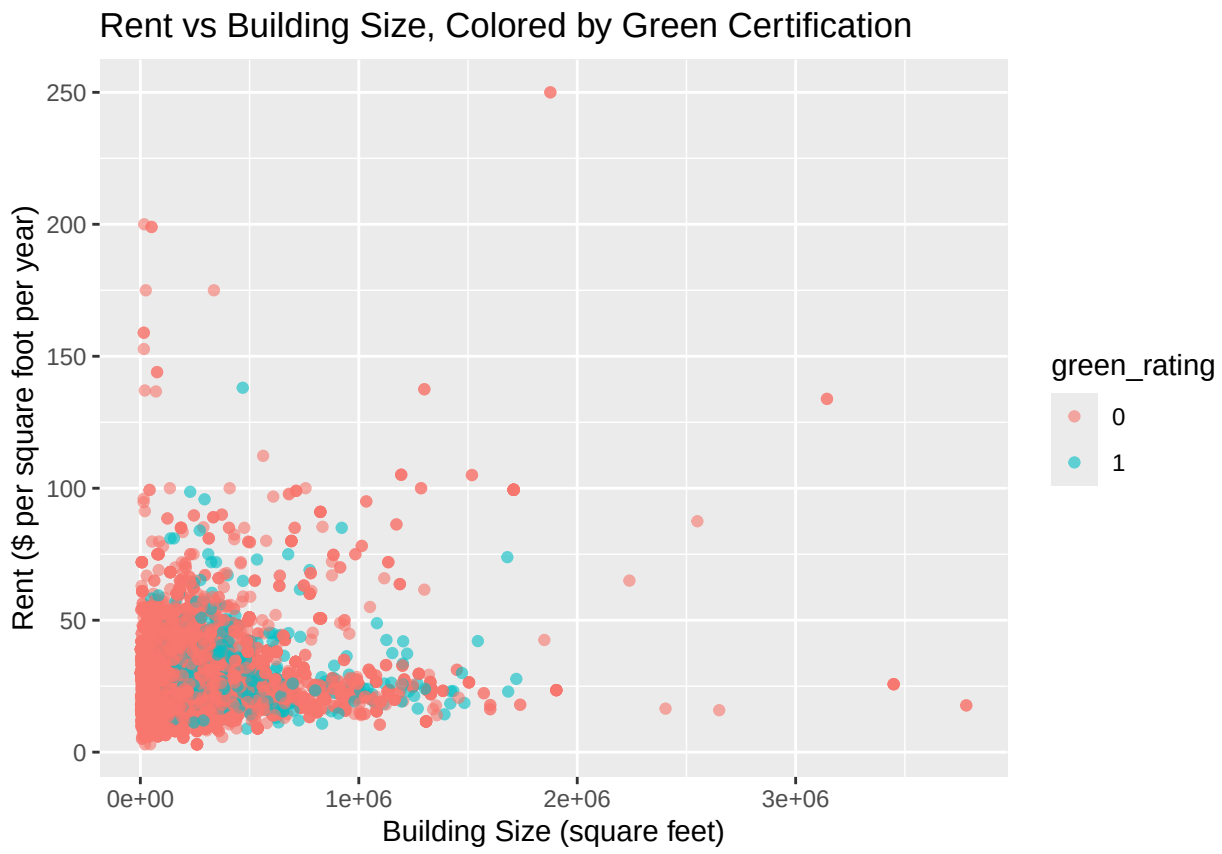

```
ggplot(data, aes(x = age, y = Rent, color = green_rating)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = FALSE, color = "black") +  # Add a black trend line
  labs(title = "Rent vs Building Age, Colored by Green Certification",
       x = "Building Age (years)",
       y = "Rent ($ per square foot per year)") +
  facet_wrap(~ green_rating)  # Facet by green certification
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

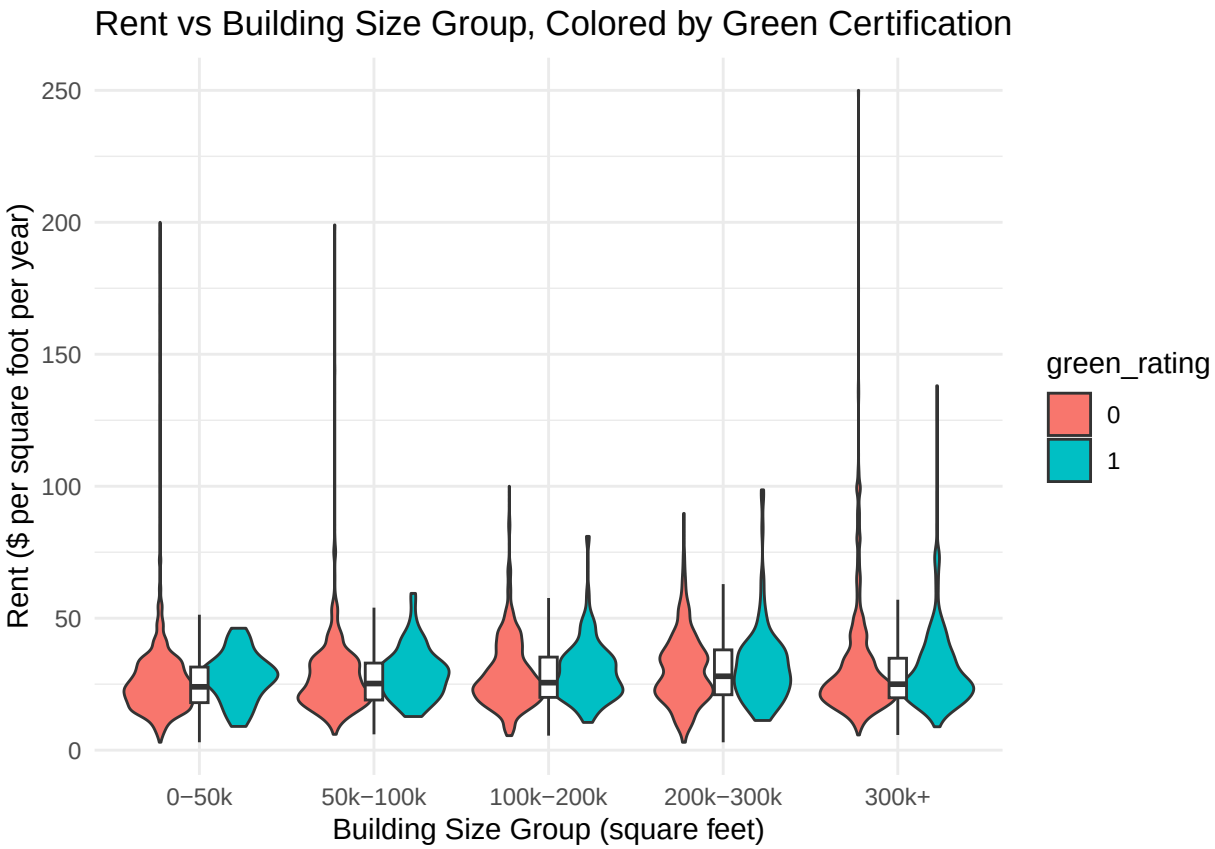
Insights - The scatterplot shows that rent generally decreases as building age increases, which is a common trend in real estate. However, green-certified buildings tend to maintain higher rents across different building ages compared to non-certified ones. The black trend line in the scatterplot reinforces this observation, indicating that green certification can help sustain higher rents even as buildings age. The faceted boxplot further supports this, showing that green certification contributes positively to rent within each age group.

Rent by Building Size

```
# Scatterplot: Rent vs Building Size colored by Green Certification
ggplot(data, aes(x = size, y = Rent, color = green_rating)) +
  geom_point(alpha = 0.6) +
  labs(title = "Rent vs Building Size, Colored by Green Certification",
       x = "Building Size (square feet)",
       y = "Rent ($ per square foot per year)")
```



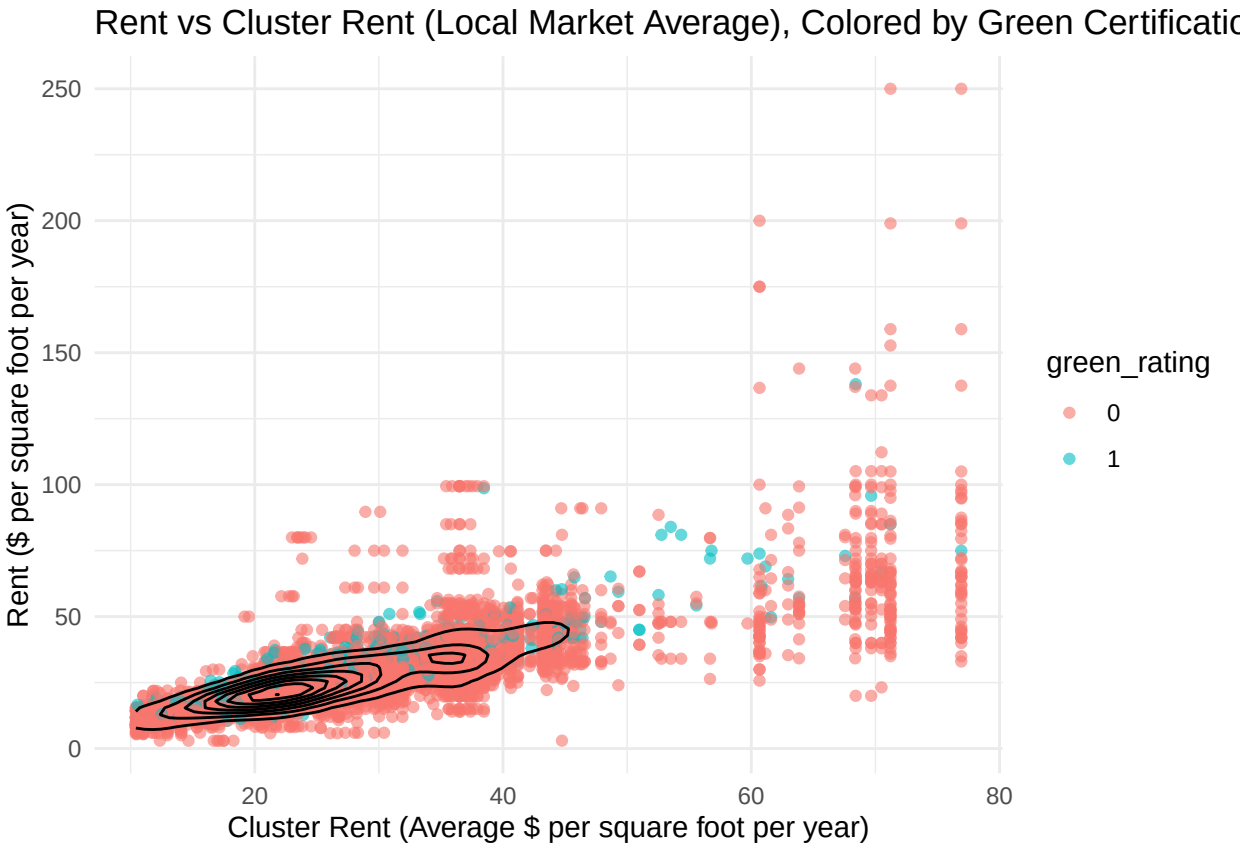

```r
library(dplyr)

# Create size bins
data <- data %>%
  mutate(size_group = cut(size,
                          breaks = c(0, 50000, 100000, 200000, 300000, Inf),
                          labels = c("0-50k", "50k-100k", "100k-200k", "200k-300k", "300k+")))

# Now plot the data using the size_group variable
ggplot(data, aes(x = size_group, y = Rent, fill = green_rating)) +
  geom_violin() +
  geom_boxplot(width = 0.1, fill = "white", outlier.shape = NA) +  # Overlay boxplot for summary stats
  labs(title = "Rent vs Building Size Group, Colored by Green Certification",
       x = "Building Size Group (square feet)",
       y = "Rent ($ per square foot per year)") +
  theme_minimal()
```



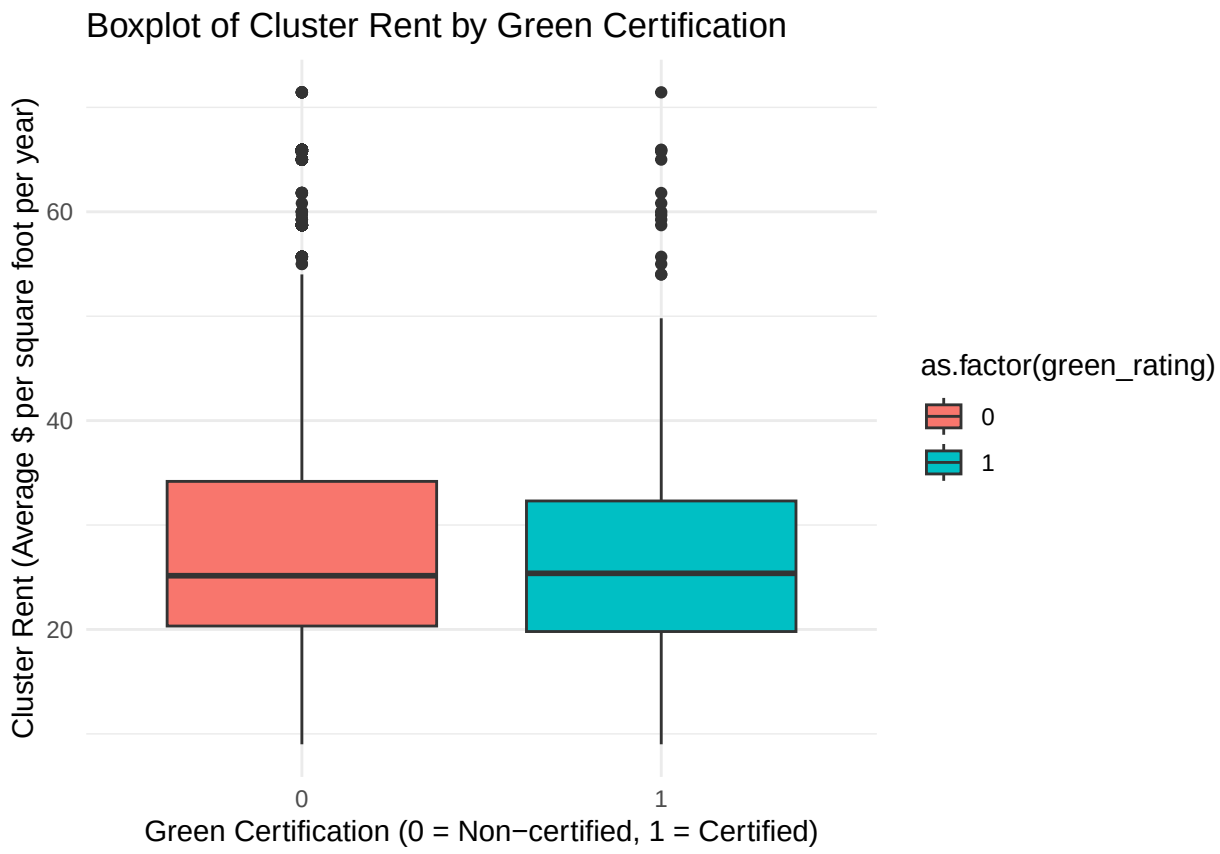
Insights - The scatterplot does not show a strong or clear relationship between building size and rent. Both small and large buildings can have high or low rents, and green-certified buildings are distributed across various sizes. The violin plot, with an overlaid boxplot, provides a detailed view of the rent distribution across different size groups. The plot suggests that building size is not a significant confounder in this analysis, and the rent premium associated with green certification is consistent across different building sizes.

Rent by Cluster Rent

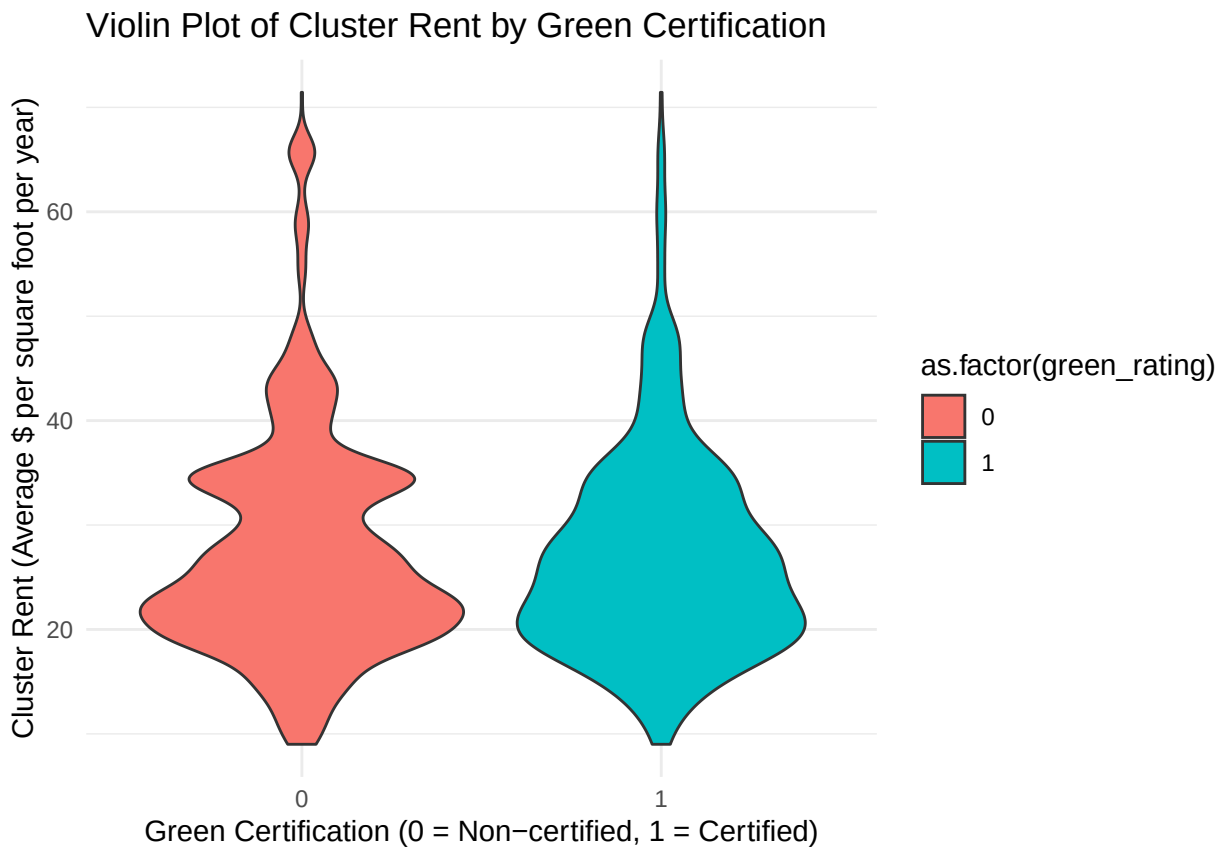
```
# Scatterplot: Rent vs Cluster Rent (local market average) colored by Green Certification
library(dplyr)

# Calculate the average rent for each cluster
data <- data %>%
  group_by(cluster) %>%
  mutate(cluster_rent_avg = mean(Rent, na.rm = TRUE))


ggplot(data, aes(x = cluster_rent_avg, y = Rent, color = green_rating)) +
  geom_point(alpha = 0.6) +
  geom_density_2d(color = "black") +  # Add black density lines for better visibility
  labs(title = "Rent vs Cluster Rent (Local Market Average), Colored by Green Certification",
       x = "Cluster Rent (Average $ per square foot per year)",
       y = "Rent ($ per square foot per year)") +
  theme_minimal()
```




```
ggplot(data, aes(x = as.factor(green_rating), y = cluster_rent, fill = as.factor(green_rating))) +
  geom_boxplot() +
  labs(title = "Boxplot of Cluster Rent by Green Certification",
       x = "Green Certification (0 = Non-certified, 1 = Certified)",
       y = "Cluster Rent (Average $ per square foot per year)") +
  theme_minimal()
```




```
ggplot(data, aes(x = as.factor(green_rating), y = cluster_rent, fill = as.factor(green_rating))) +
  geom_violin() +
  labs(title = "Violin Plot of Cluster Rent by Green Certification",
       x = "Green Certification (0 = Non-certified, 1 = Certified)",
       y = "Cluster Rent (Average $ per square foot per year)") +
  theme_minimal()
```

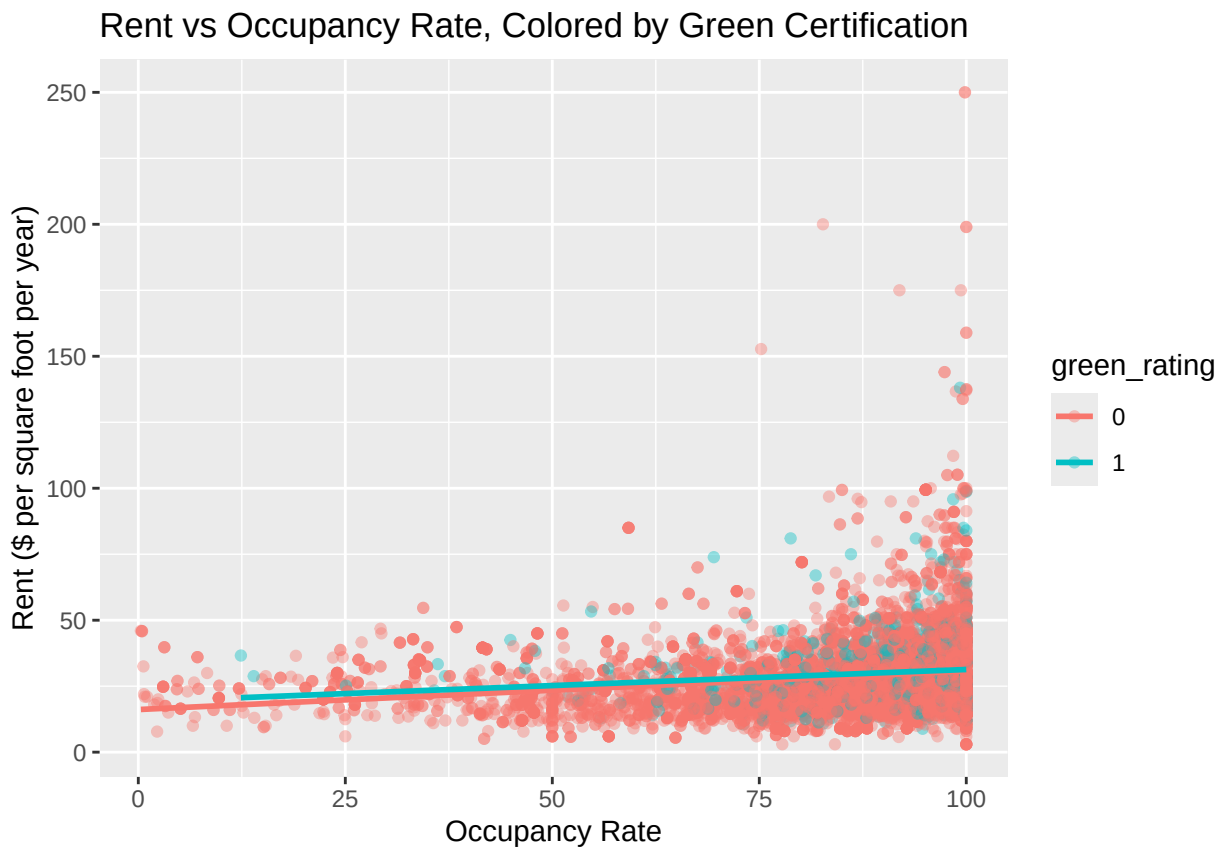


Insights - Within these clusters, we have only one green building in each cluster. The scatterplot and density lines plotted above indicate that green buildings tend to have slightly higher rents than the surrounding buildings. This suggests that while location is important, green certification still contributes to higher rents, independent of location.

Rent by Occupancy Rate

```
# Scatterplot: Rent vs Occupancy Rate colored by Green Certification
ggplot(data, aes(x = leasing_rate, y = Rent, color = green_rating)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Rent vs Occupancy Rate, Colored by Green Certification",
       x = "Occupancy Rate",
       y = "Rent ($ per square foot per year)")
```

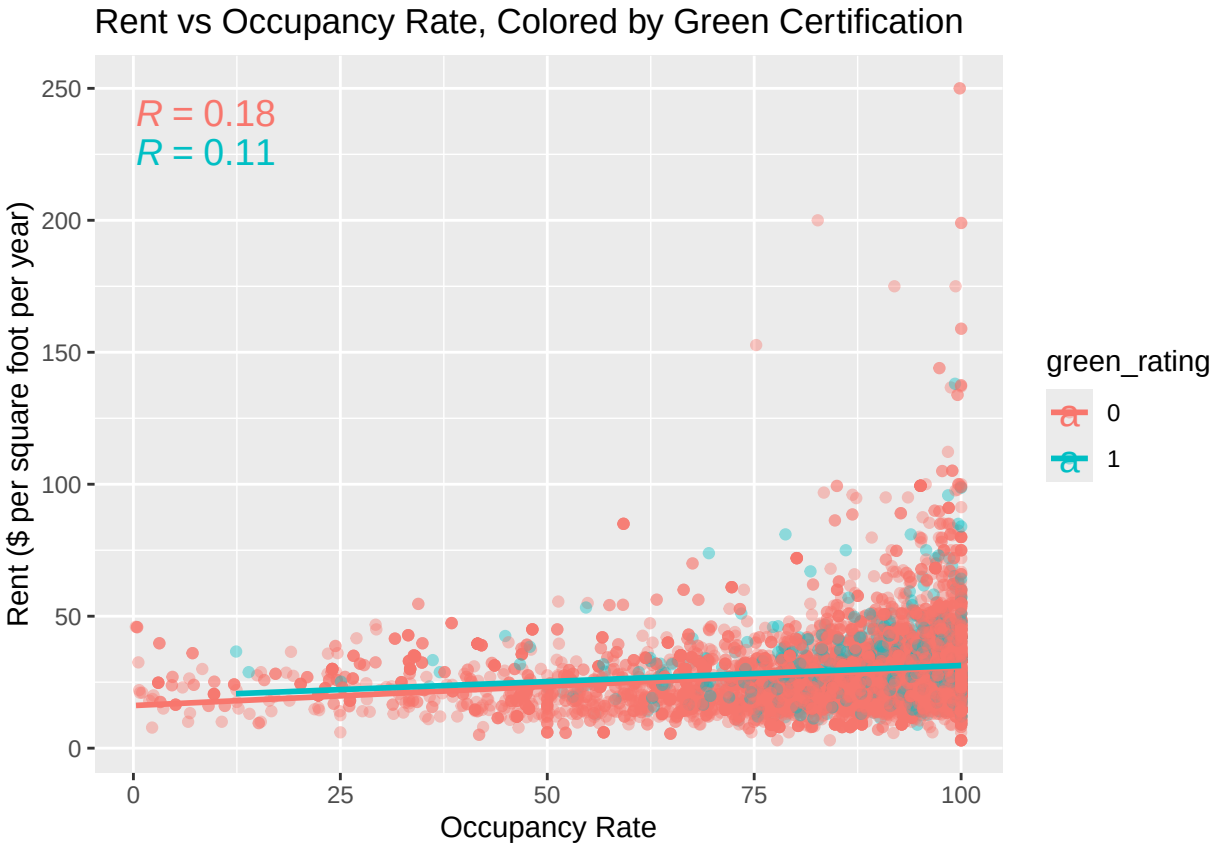
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
library(ggpubr)

ggplot(data, aes(x = leasing_rate, y = Rent, color = green_rating)) +
  geom_point(alpha = 0.4, position = position_jitter(width = 0.02, height = 0)) +  # Add jitter
  geom_smooth(method = "lm", se = FALSE) +
  stat_cor(method = "pearson", aes(label = ..r.label..), size = 5) +  # Add correlation coefficient
  labs(title = "Rent vs Occupancy Rate, Colored by Green Certification",
       x = "Occupancy Rate",
       y = "Rent ($ per square foot per year)")
```

```
## Warning: The dot-dot notation (‘..r.label..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(r.label)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



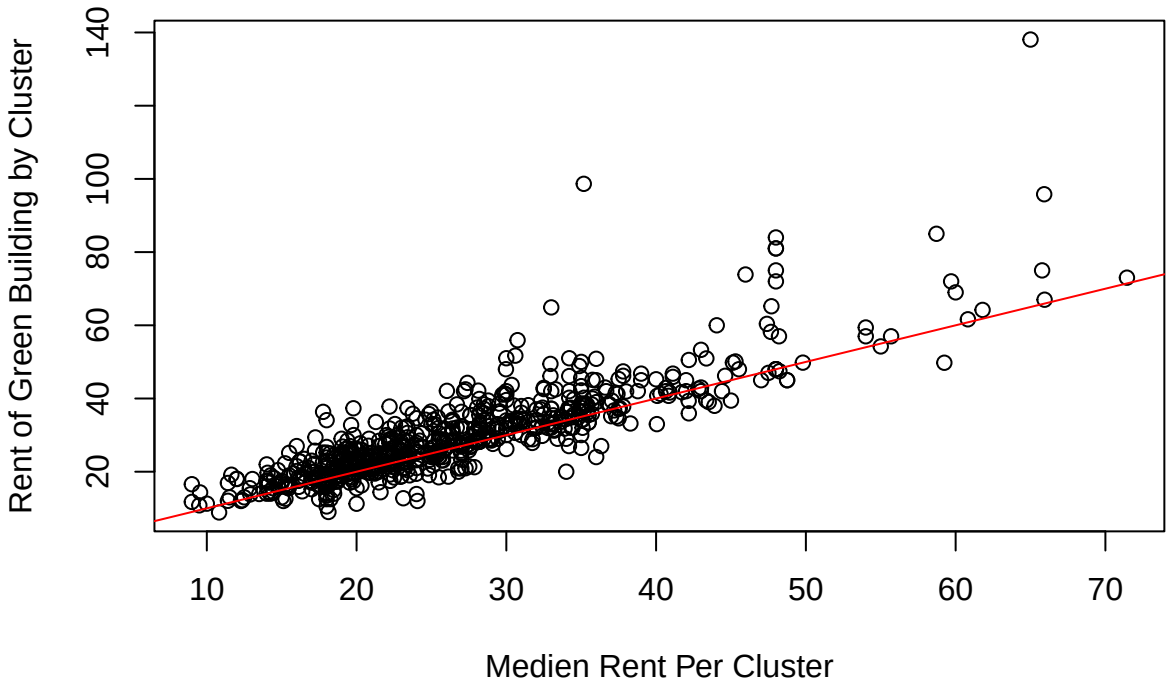
Insights - The scatterplot with a linear trend line shows that there is no significant difference in occupancy rates between green-certified and non-certified buildings. Both types of buildings track closely in terms of occupancy rate, and green certification does not seem to significantly impact occupancy rates. Therefore, the higher rents associated with green certification are not due to higher occupancy rates but are likely influenced by other factors.

Some other summary stats

```
library(tidyr)
library(dplyr)

median_rent_by_class <- data %>%
  group_by(green_rating, class) %>%
  summarize(median_rent = median(Rent, na.rm = TRUE), .groups = 'drop') %>%
  pivot_wider(names_from = class, values_from = median_rent)

# Display the median rent by green rating and class
print(median_rent_by_class)
```

```
## # A tibble: 2 x 4
##   green_rating `Class A` `Class B` `Class C`
##   <fct>            <dbl>     <dbl>     <dbl>
## 1 0                 28.2      24        22.1
## 2 1                 28.4      25.2      32
```

```r
data2 <- filter(data, green_rating == 1)

plot(x = data2$cluster_rent, y = data2$Rent, xlab = 'Medien Rent Per Cluster', ylab = 'Rent of Green Bu
abline(a=0, b=1, col = 'red')
```




```
## integer(0)
```

Conclusion and scope of further investigation - Green certification does have a positive impact on rent (as suggested by the Excel guru), but its effect is modest and varies depending on on the certain underlying factors like building class, age, and location. Building size and occupancy rate do not appear to be significant confounders in this analysis. Therefore, we need to do further analysis on these underlying factors.

##

Objective recap - Determine whether investing in a green building certification for a new 15-story, 250,000 square foot building in Austin is economically viable.

Approach - The building is expected to have certain characteristics (e.g., amenities, stories, size) that should be compared to similar buildings in the data set to assess the potential financial impact of green certification.

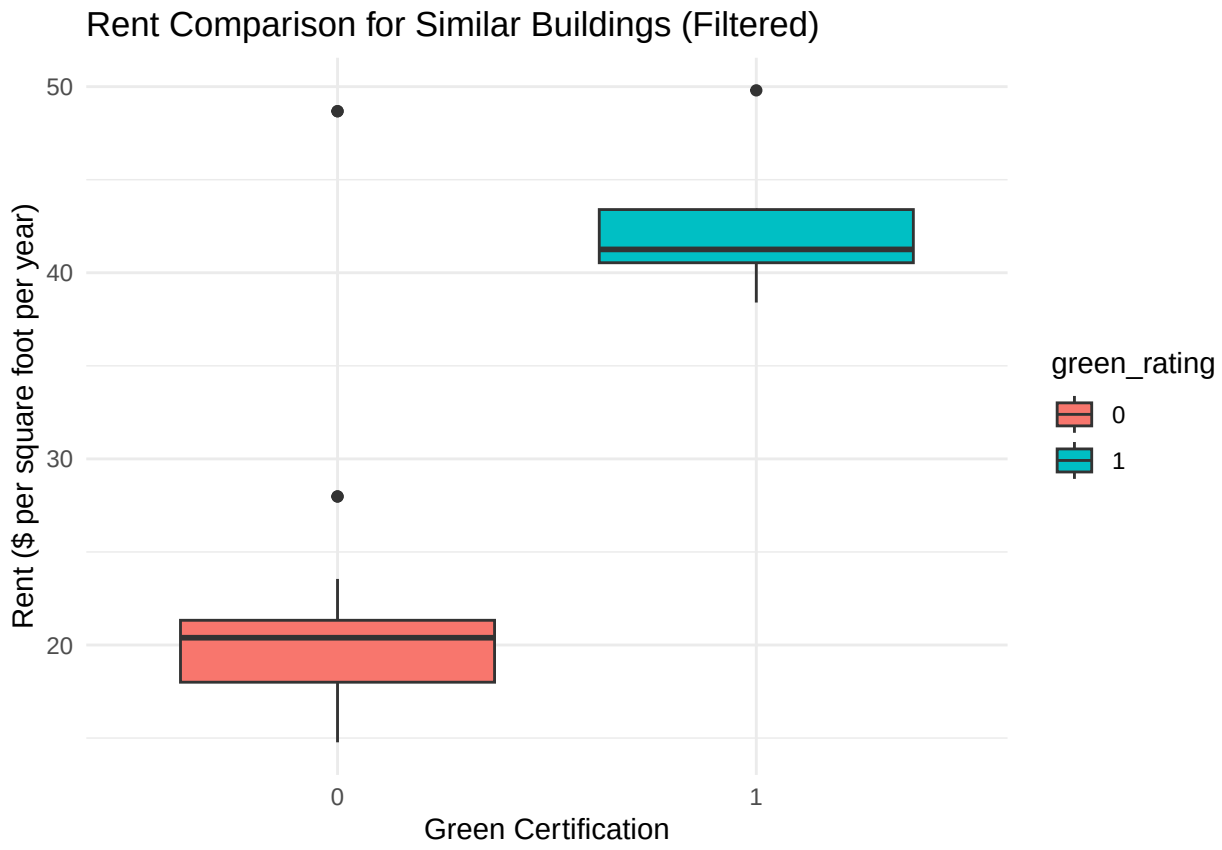
#Step 1 - Filtering for similar buildings to match the specific criteria (15 stories, presence of amenities, and 250,000 sq ft in size) to isolate a subset of buildings that are most comparable to the one under consideration.

#Step 2 - Analyzing leasing rates to ensure that the buildings in the subset have comparable occupancy levels, which is a key factor in determining rent levels and overall financial performance.

i.e. Stories = 15, and Size = 250,000

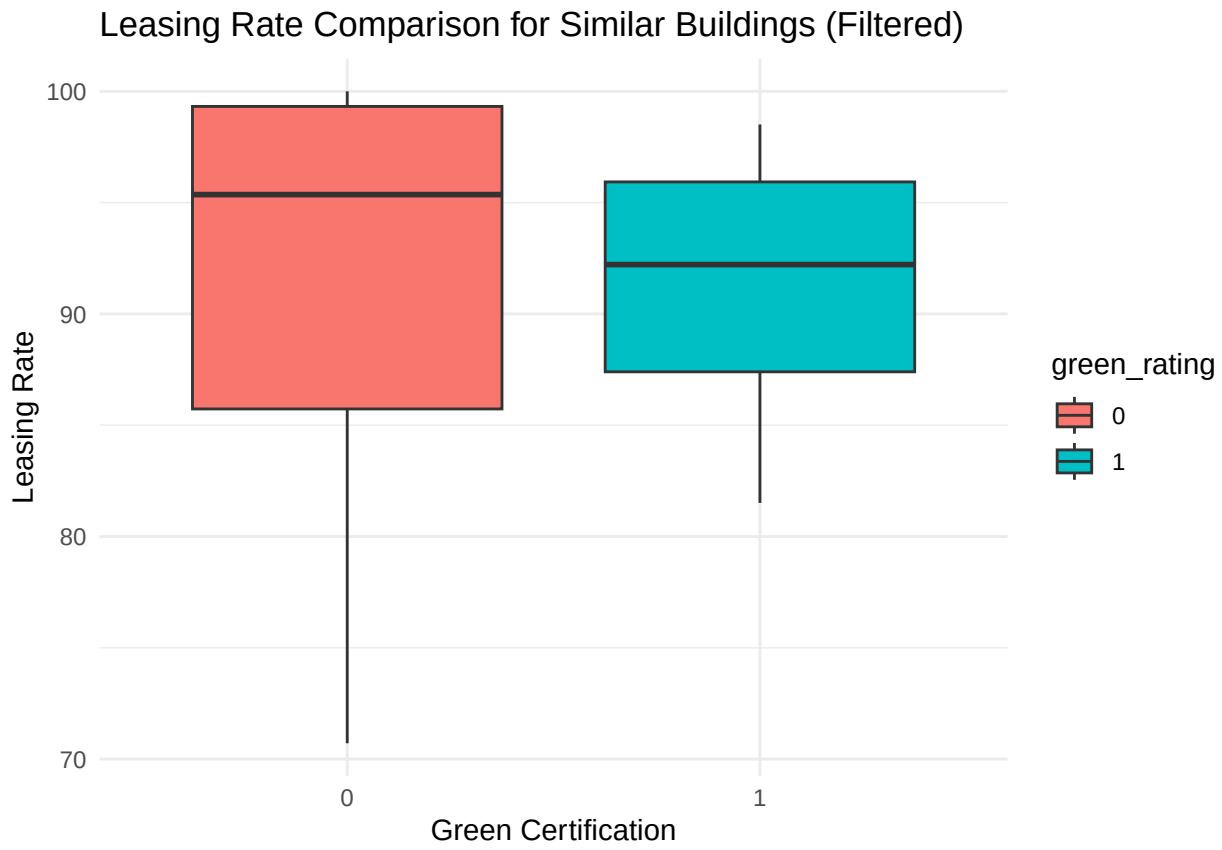
This does seem like it has amenities based on the description so I will compare to buildings with amenities

```
library(ggplot2)


# Filter buildings based on size, stories, and amenities
final_data <- data %>%
  filter(stories == 15, size >= 225000 & size <= 275000)

# Buildings with leasing rates below the 10th percentile are already removed

# Create a boxplot to compare rents
ggplot(final_data, aes(x = green_rating, y = Rent, fill = green_rating)) +
  geom_boxplot() +
  labs(title = "Rent Comparison for Similar Buildings (Filtered)",
       x = "Green Certification",
       y = "Rent ($ per square foot per year)") +
  theme_minimal()
```



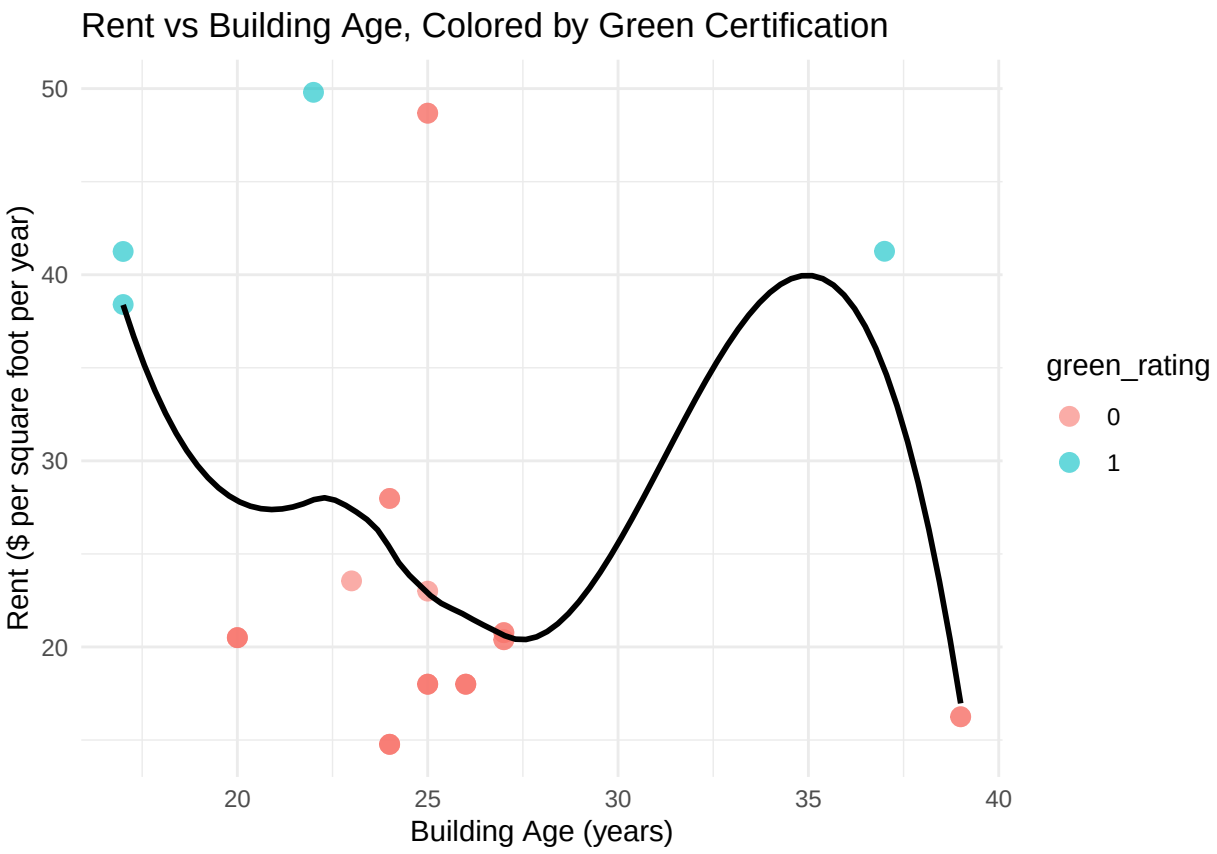

```
# Visualize leasing rates across green certification
ggplot(final_data, aes(x = green_rating, y = leasing_rate, fill = green_rating)) +
  geom_boxplot() +
  labs(title = "Leasing Rate Comparison for Similar Buildings (Filtered)",
       x = "Green Certification",
```

```
       y = "Leasing Rate") +
  theme_minimal()
```



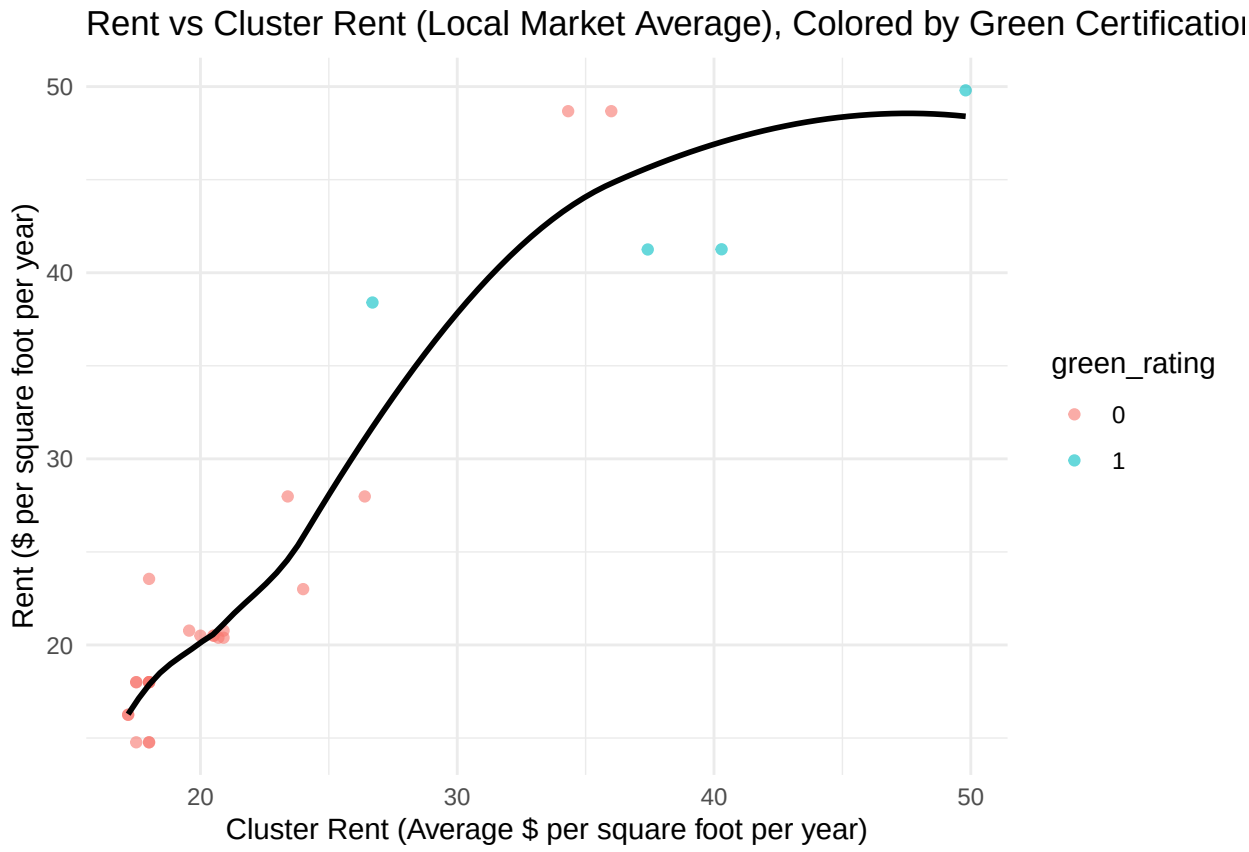

```
# Explore the Impact of Building Age on Rent:
ggplot(final_data, aes(x = age, y = Rent, color = green_rating)) +
  geom_point(alpha = 0.6, size = 3) +
  geom_smooth(method = "loess", se = FALSE, color = "black") +
  labs(title = "Rent vs Building Age, Colored by Green Certification",
       x = "Building Age (years)",
       y = "Rent ($ per square foot per year)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Compare Cluster Rent (Local Market Average) for Green and Non-Green Buildings:
ggplot(final_data, aes(x = cluster_rent, y = Rent, color = green_rating)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = FALSE, color = "black") +
  labs(title = "Rent vs Cluster Rent (Local Market Average), Colored by Green Certification",
       x = "Cluster Rent (Average $ per square foot per year)",
       y = "Rent ($ per square foot per year)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Since the sample size is small, we will check the summary statistics

```
summary_stats <- final_data %>%
  group_by(green_rating) %>%
  summarize(
    Mean_Rent = mean(Rent, na.rm = TRUE),
    Median_Rent = median(Rent, na.rm = TRUE),
    Mean_Leasing_Rate = mean(leasing_rate, na.rm = TRUE),
    Median_Leasing_Rate = median(leasing_rate, na.rm = TRUE)
  )

print(summary_stats)
```

```
## # A tibble: 2 x 5
##   green_rating Mean_Rent Median_Rent Mean_Leasing_Rate Median_Leasing_Rate
##   <fct>            <dbl>       <dbl>             <dbl>               <dbl>
## 1 0                 22.0        20.4              90.7                95.4
## 2 1                 42.7        41.3              91.1                92.2
```

Final Recommendation

```
leasing_rate_by_class_and_green <- data %>%
  group_by(class, green_rating) %>%
  summarize(mean_leasing_rate = mean(leasing_rate, na.rm = TRUE),
```

```
            median_leasing_rate = median(leasing_rate, na.rm = TRUE),
            .groups = 'drop')
# Display the results
print(leasing_rate_by_class_and_green)
```

```
## # A tibble: 6 x 4
##   class   green_rating mean_leasing_rate median_leasing_rate
##   <chr>   <fct>                   <dbl>               <dbl>
## 1 Class A 0                        88.0                92.7
## 2 Class A 1                        90.1                93.6
## 3 Class B 0                        82.8                88.5
## 4 Class B 1                        86.5                89.7
## 5 Class C 0                        76.6                83.3
## 6 Class C 1                        88.2                91.3
```

Based on our analysis of the available data, which includes a small sample of 28 buildings, we observed that green-certified buildings tend to command higher rents compared to non-certified ones. However, due to the limited sample size, these findings should be interpreted with caution. To ensure a well-informed decision, we recommend further research and data collection, possibly expanding the dataset to include similar markets. If these trends hold in a larger sample, pursuing green certification could be a viable option to enhance the building's marketability and financial performance

As of now, based on the entire sample, we would not go with the procuring the green certification for the building. While the green buildings do have higher rent, it is only 92 cents higher per square foot (i.e. 0.9% added income per square foot). We base our calculations on the assumption that the new building would have been a Class A building based on the description of the quality.

We are incurring an added expense of 5Mn to make half a percent more income over the life of the building (under the assumption that the rents remain stable throughout the life of the building). While we did like the idea of creating a green building from environmental point of view, there is limited benefit from an economical perspective. We would like more data regarding the buidling like the number of predicted days of using heating/cooling e.g. In Austin, we definitely use more cooling in the summer which is cheaper for green buildings according to the description but this is not included in the potential building. We would not proceed with the certification based on the limited data about the new building and perform a more detailed analysis.