Exercise 1

Stephenson Gokingco, Akash Thakkar, Caroline Hao, James Cornejo

2/14/2020

1) Flight Departure Delays from Austin-Bergstrom International Airport

Using 2008 data on flights passing through Austin-Bergstrom International Airport (AUS), our team wanted to identify trends in average departure delays for flights leaving AUS based on time of departure and day of the week. To do this, we chose to create a bar graph of average departure delays in minutes on the vertical-axis and time of departure on the horizontal-axis.

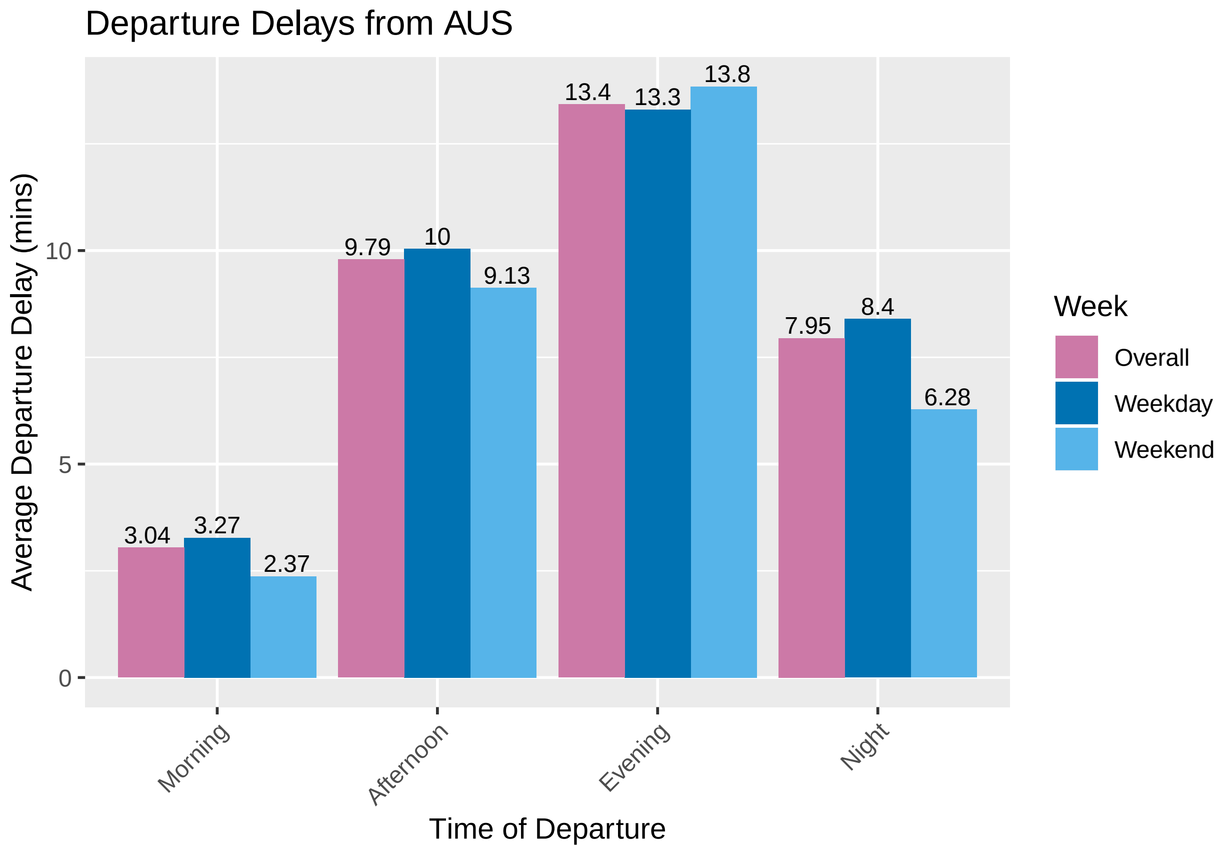
Flights leaving from 5:00 AM to 11:59 AM were classified as morning departures. Flights leaving from 12:00 PM to 4:59 PM were classified as afternoon departures. Flights leaving from 5:00 PM to 8:59 PM were classified as evening departures Lastly, flights leaving from 9:00 PM to 4:49 AM were classified as night departures.

Furthermore, there are three bars for each departure time category. The first bar plots the average delay for all of the flights leaving AUS. The second and third bars plot the average delay for weekday and weekend flights respectively.

Libraries and Loading Dataset:

To create the plot, we used the 'ABIA.csv' file and the ggplot2, dplyr, and tidyverse libraries.

```r
library(ggplot2)
library(dplyr)
library(tidyverse)
abia <- read.csv(params$abia)
```



```
## # A tibble: 4 x 4
##   timeofdep count.weekday count.weekend count.all
##   <fct>             <int>         <int>     <int>
## 1 Morning           17058          5875     22933
## 2 Afternoon         11468          4291     15759
## 3 Evening            8104          2386     10490
## 4 Night               347            94       441
```

From the bar plot, a couple trends emerge that can be useful for passengers flying out from AUS. First, morning flights have the shortest average departure delay of around 3 minutes. Similarly, evening flights have the longest average departure delay of about 13 minutes. Second, weekend flights tend to have shorter delays than weekday flights, with the exeption of evening departures.

To accompany the bar plot, we have also included a table of the number of flights by weekday/weekend status and time of departure. Because the vertical-axis of the bar graph depicts the average departure delay, it is good to be aware of the number of flights used to compute the averages for each bar. Morning flights are by far the most common, followed by afternoon, and then evening departures. Night flights are rather rare and made up less than 1% of all flights departing AUS in 2008.

Overall, the trends discussed above can be especially useful to know for people who travel often and on tight schedules. A couple of minutes can make the difference of what side of the door of a connecting flight you find yourself after the final-boarding-call. Furthermore, if a traveller needs to book a flight on short notice, she likely will be able to find a morning option, but unable to find a night departure.

2) Creatinine Clearance Rate vs. Age

Given data regarding a patient's age and their respective creatine clearance rate, we wanted to answer the following:

1. Average creatinine clearance rate for a 55-year-old.
2. Creatinine clearance rate of change with age.
3. Given two individuals, who is healthier for their age?

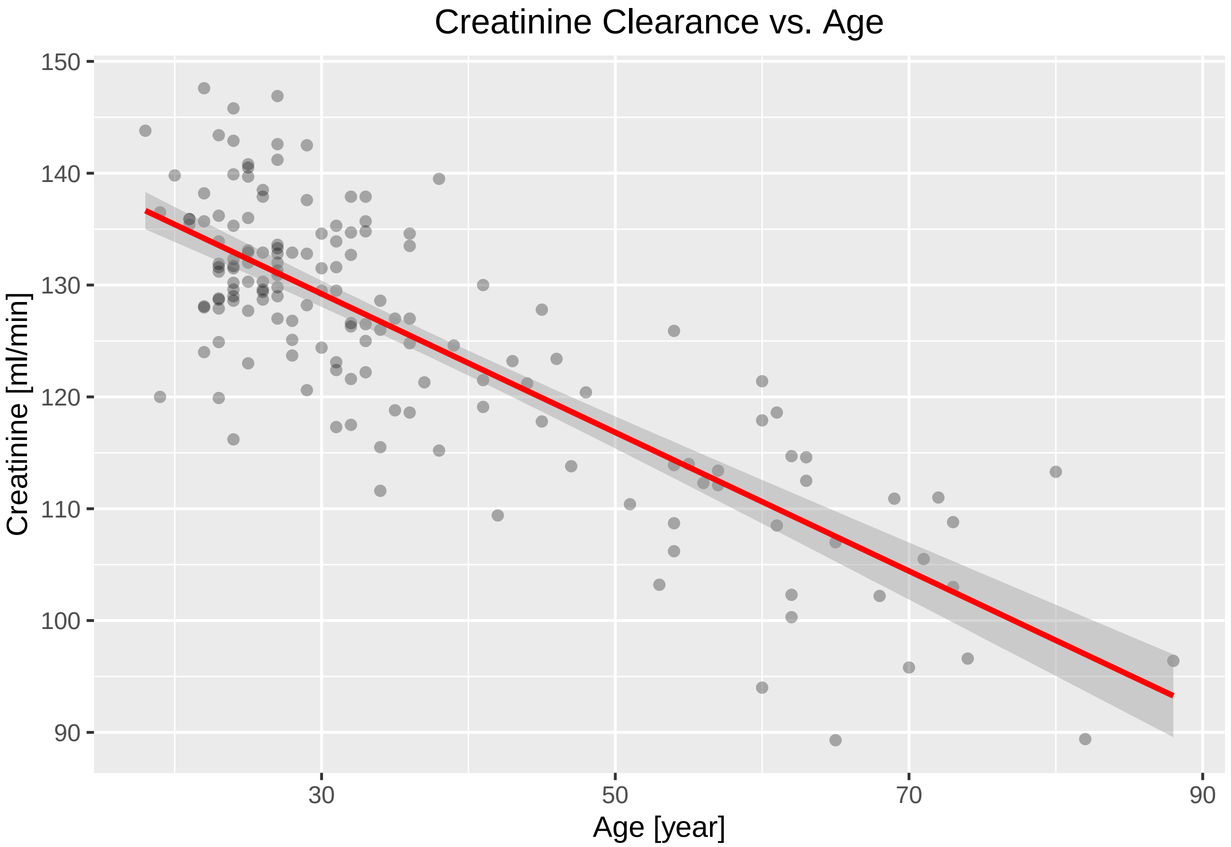
- 40-year-old | CCR 135
- 60-year-old | CCR 112

Approach

For the analysis, we used a 'creatinine.csv' file and the ggplot2 library.

```
library(ggplot2)
creatinine <- read.csv(params$creatinine)
```

We fit a linear regression model on data relating creatinine clearances rates and age. We used this model to answer whether or not an individual is healthy compared to the average clearance rate at their respective age groups and answer the various questions stated above.



From this plot, it is clear to see the decrease in the CCR as individuals get older.

Results:

1. From the linear regression model, we were able to calculate that average clearance rate (ml/min) for a 55-year-old is:

```
## [1] 113.723
```

2. In addition, we found that the average creatinine clearance rate of change (ml/min/year) is:

```
## [1] -0.6198159
```

3. Comparing a 40-year-old w/ 135 CCR and a 60-year-old with 112 CCR:

- To compare these values, I calculated the average CCR for 40-year-olds and 60-year-olds respectively. With these values, I compared the average with each of their respective CCR rates and calculated the percentage difference between them.

40-year-old CCR (ml/min)

```
## [1] 9.738003
```

60-year-old (ml/min)

```
## [1] 1.243885
```

In order to distinguish who is healthier, we understood that a higher creatinine clearance rate is better.

Since the fourty year old is around **9.7%** higher than the average and the sixty year old is around **1.2%** higher than the average at their each relative age. We concluded that the fourty old is relatively healthier in their age group.

3) Green Buildings Rent Analysis

Approach

We set out to determine whether our hypothetical analyst's assessment that a builidngs green rating alone could increase the rent charged by a substantial amount. The problem we identified with the analysis presented is that the analyst simply compared a group of green buildings to a group of non-green buildings without attention to whether there might be other systematic differences between the two groups. It seemed intuitive to us that green buildings would more frequently have other factors that drove higher prices compared to non-green buildings. For example, green buildings would likely be newer an average than non-green buildings considering environmental concerns are a fairly recent phenomenon.

In order to test this hypothesis, we set out to visualize systematic differences between green and non-green buildings in the greenbuildings dataset. We looked at the variables available and hypothesized which ones could potentially be counfounding variables for higher price. Then, we visualized the average values each of them across green and non-green rated buildings. We have included the factors with the clearest differences: number of stories, building age, and building classification type.

Results

1) Number of Stories

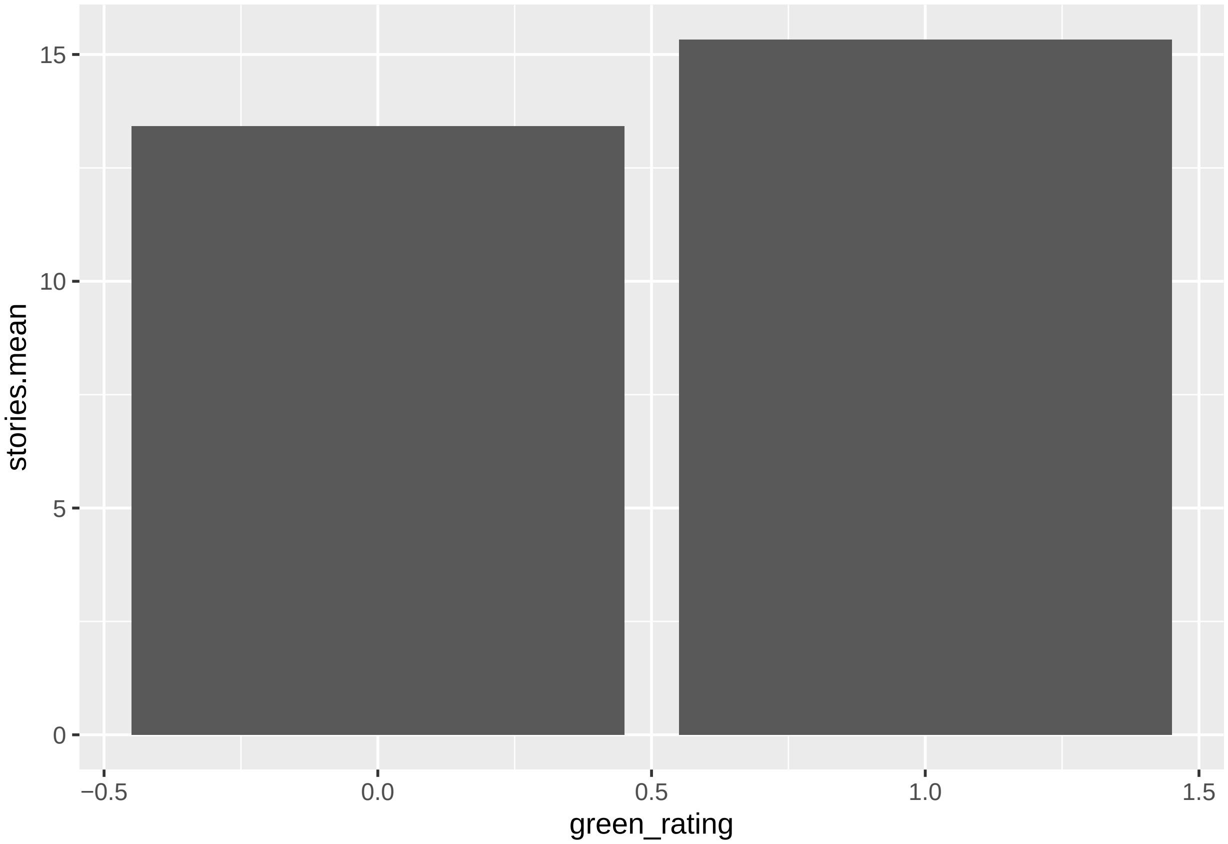


Figure 1: There seems to be a higher proportion of taller green buildings (mean ~15 stories) compared to non-green buildings (mean ~13 stories). Taller buildings can charge higher rent rates because they achieve better ventilation and light.

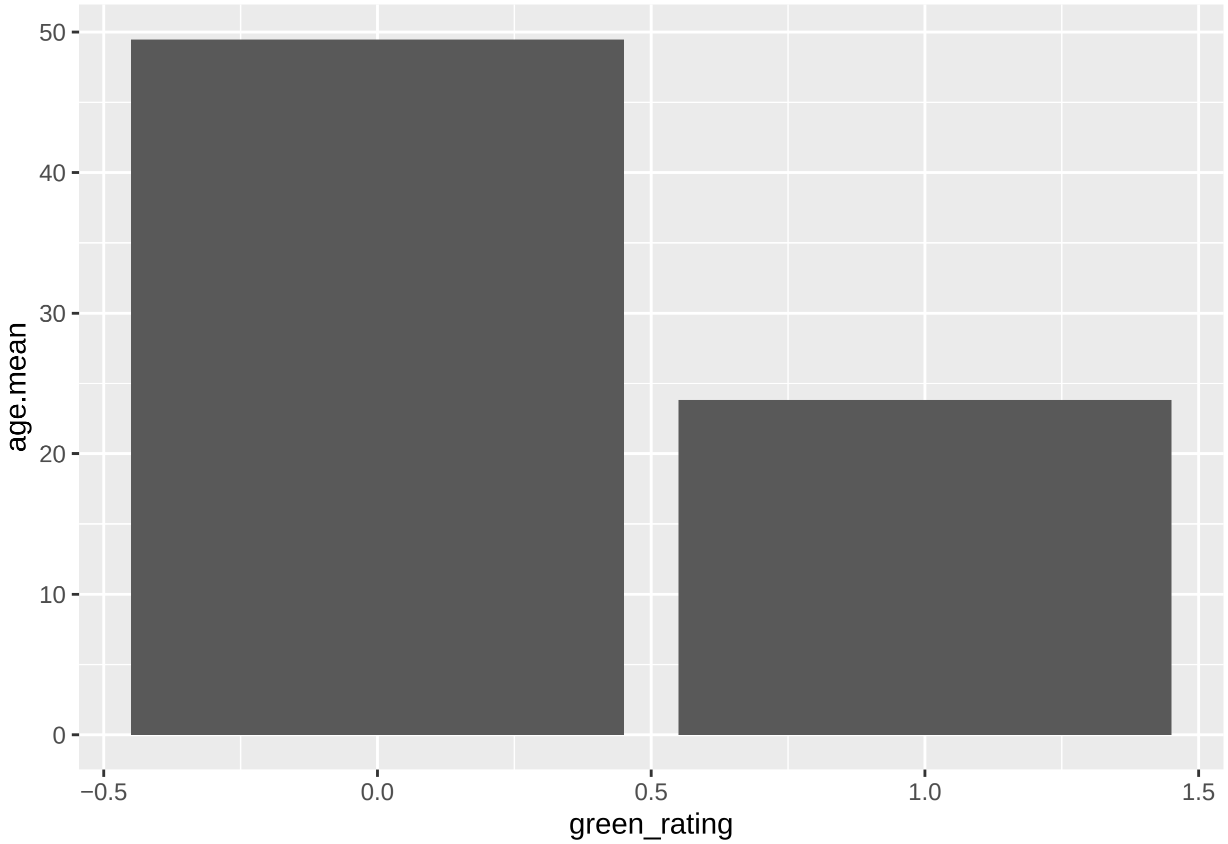


Figure 2: Green buildings (mean ~25 years) are substantially newer than non-green buildings (mean ~50 years). Newer buildings tend to be more expensive than older buildings.

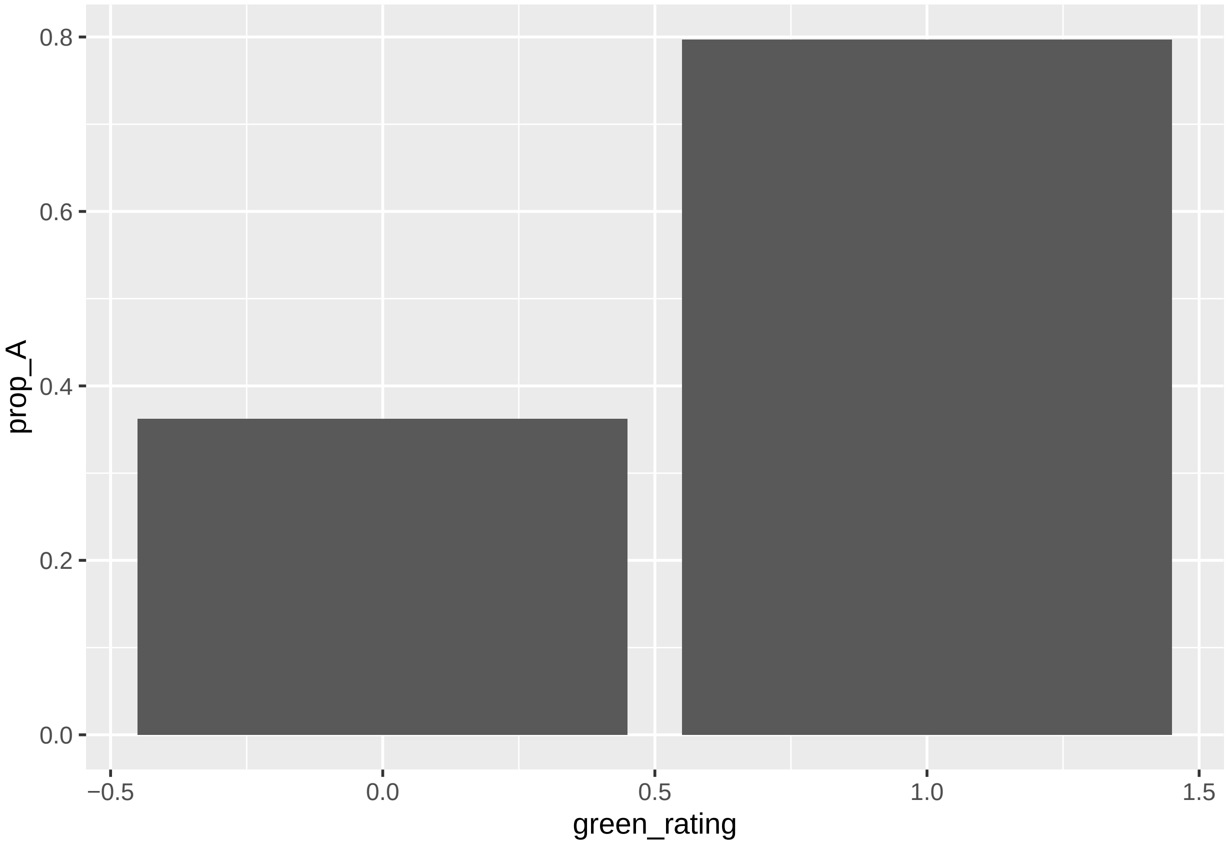


Figure 3: Green buildings are disproportionately class A buildings (~80%) compared to non-green buildings (~35%). Class A buildings are generally the highest class properties on the market and command higher rent prices.

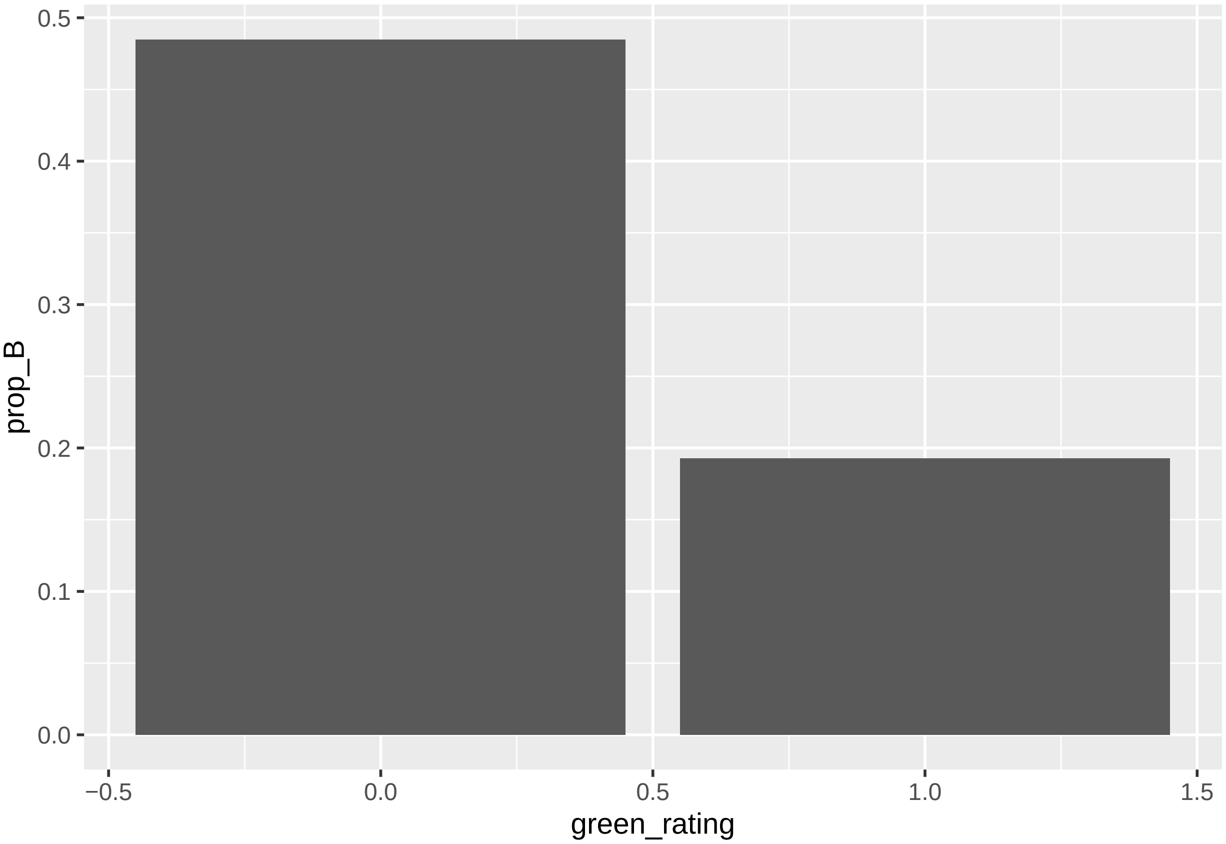


Figure 4: Non-green buildings are disproportionately class B buildings (~48%) compared to green buildings (~19%). Class B buildings are generally lower class properties compared to type A buildings and command lower rent prices.

Conclusion

From this analysis, we determined that there are substantial systematic differences between the green and non-green buildings in the dataset. Green buildings tend to be higher in stories, newer, and disproportionately Class A compared to non-green buildings that tend to have fewer stories, older and disproportionately Class B. These systematic differences likely account for the large differences in rent price found between the two groups. These confounding variables make a direct comparison between green and non-green buildings in this dataset useless in determining the direct impact of green rating on rent price. In order to conduct a more rigorous analysis, the green and non-green comparison groups should be matched on all confounding variables so that we can determine the impact of the green rating on rent price specifically.

4) Milk and Maximizing Net Profit

We are given a dataset, milk.csv, which contains a random sample of daily sales figures for a small neighborhood grocery store of cartons of milk. The "price" column gives the price at which the milk was sold that day; the "sales" column says how many units were sold that day.

Let's say that the store's wholesale cost of milk is c dollars per carton. We want to maximize profit, and so we want to know how much to charge for a carton of milk in light of the information on supply and demand conveyed by the data. The following report explains our thought process, and we will express our final answer as an equation in terms of c.

Suppose that the per-unit cost c is \$1. What price should we charge, and how much net profit do we expect to make at this price?

To give a better picture, here's the first 6 lines of the data in the dataset:

```
##   price sales
## 1  3.48    15
## 2  3.12    11
## 3  2.95    21
## 4  2.68    30
## 5  3.62    11
## 6  4.32    11
```

Say that the per-unit price charged is P, and the quantity of units sold is Q. The equation that expresses net profit N in terms of both Q and P (and the per-unit cost, c) is the following:

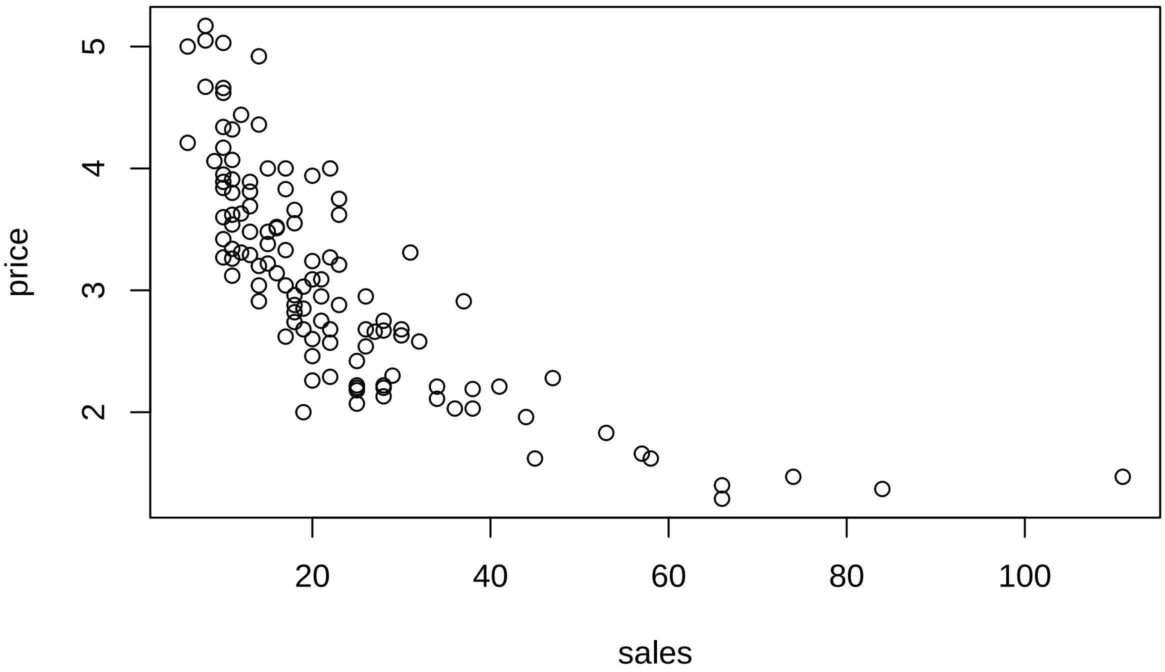
$$N = (P - c)Q$$

To naively take the first derivative with respect to P and set it equal to zero would be a mistake because quantity of milk sold is a function of the price it's set at. In other words, $Q = f(P)$. Economists use this equation to relate quantity and price:

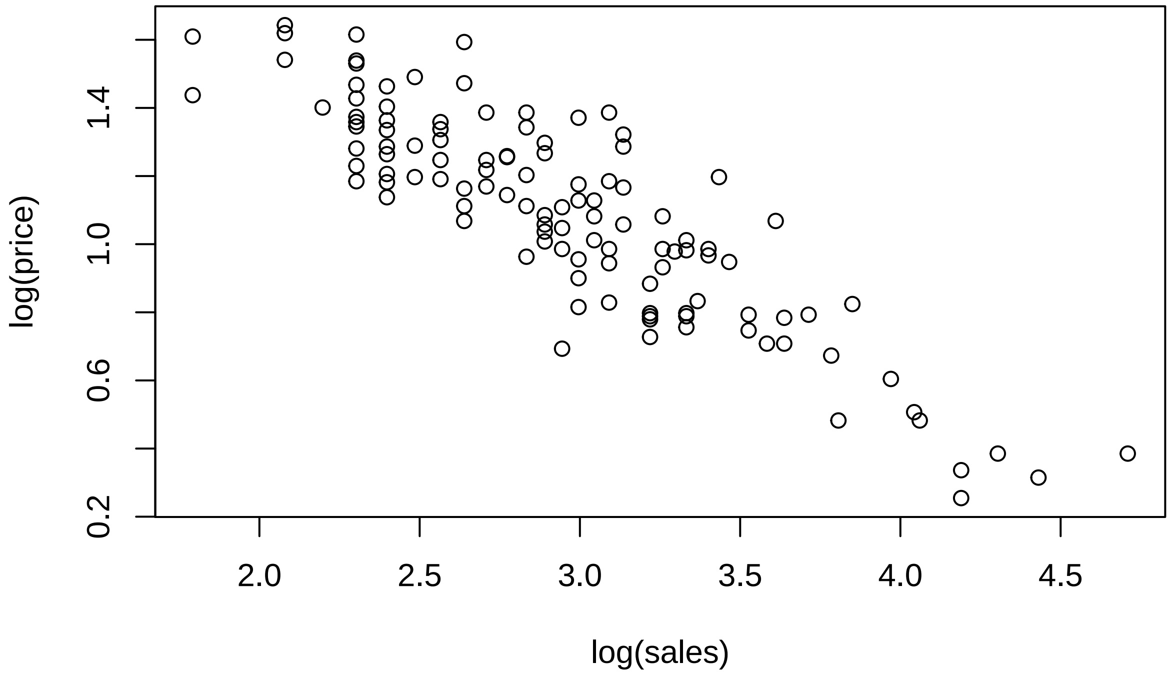
$$Q = \alpha P^{\beta}$$

where β is the price elasticity of demand. We will be conducting a regression of the data to estimate this value.

Plotting an initial, simple scatterplot of the data, we can see that there is some sort of power law relationship between price and quantity.



By taking the *log* of both variables and plotting a scatterplot, we get a relationship that looks more linear.



We get the following coefficients from a linear regression of this modified scatterplot:

```
## (Intercept)  log(price)
##    4.720604   -1.618578
```

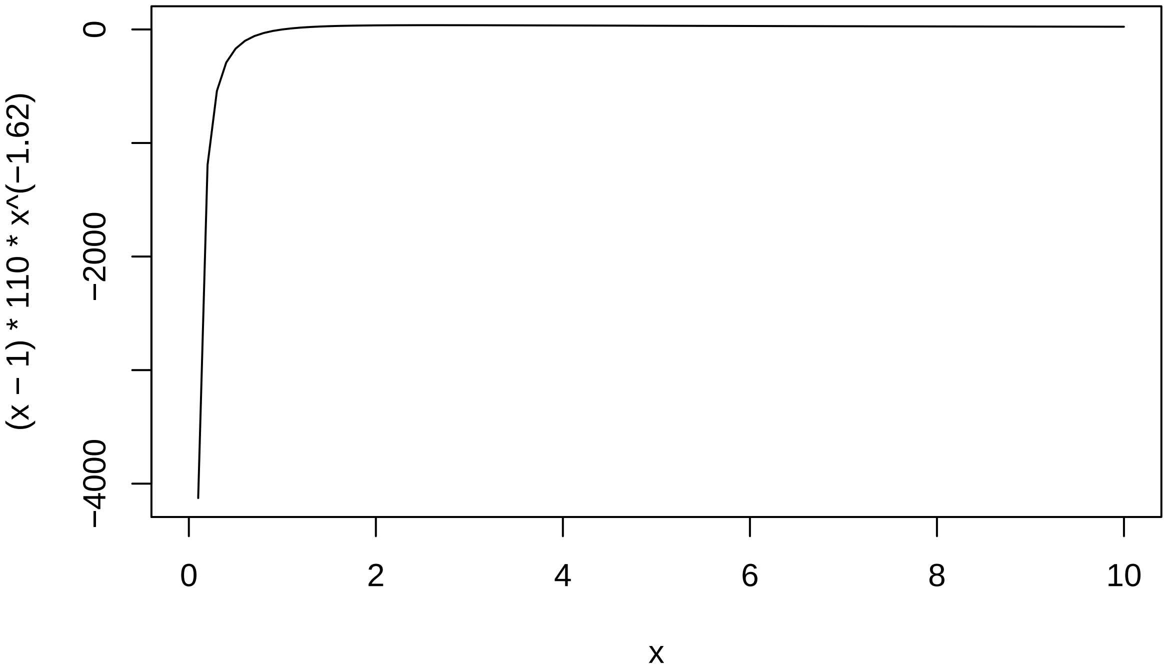
Plugging these values into our mathematical equation and then performing some algebra, we get the following:

$$log(Q) = 4.7 - 1.62log(P)$$
$$Q = e^{4.7}P^{-1.62}$$

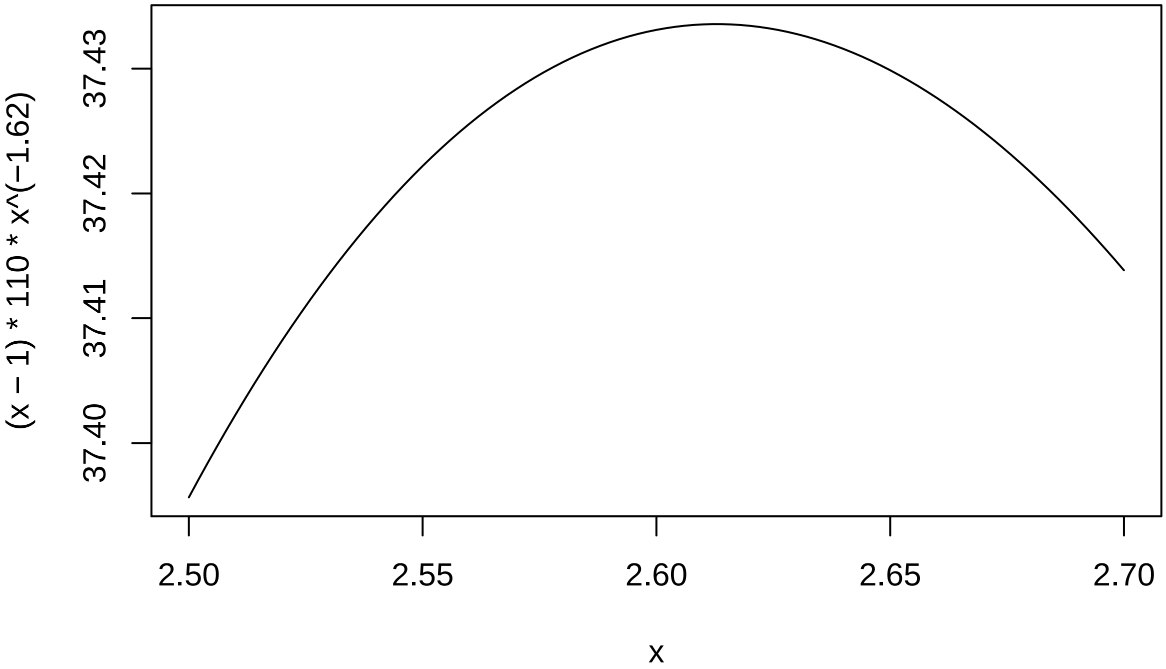
Now that we have $f(P)$, we plug this into the original equation, and we now have:

$$N = (P - 1)110P^{-1.62}$$

From this equation, we get the following mathematical curve:



We do some zooming in and the tried and true "plot and point" method to find the maximizing P.



And so, $P^* = \$2.61$, which is the price we should charge to maximize net profit. The respective $Q^* = 23.24$. The expected net profit from this price should be \$37.43.