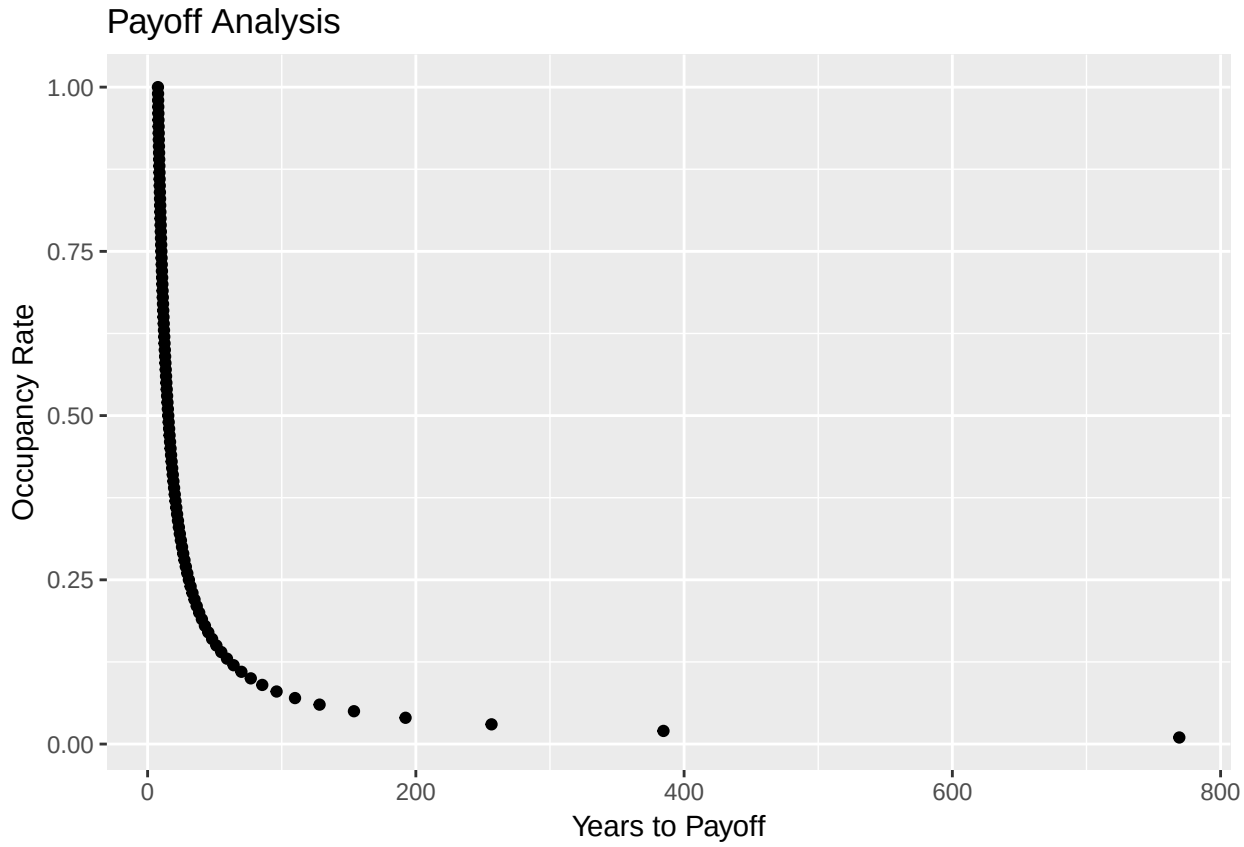
# Data Mining and Statistical Learning: Exercise 1

*Frank Chou*

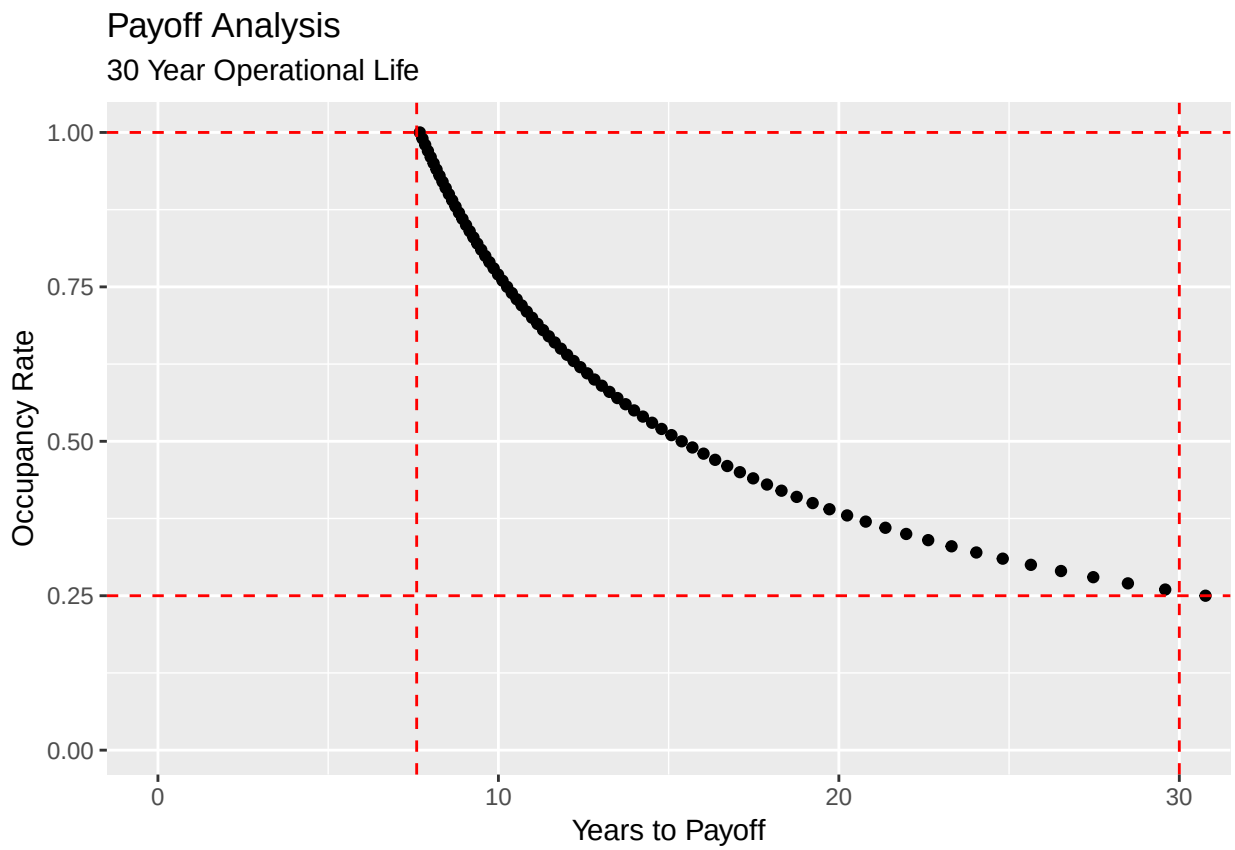
*February 9, 2019*

**Question 1**

## The Analyst's Analysis




A rudimentary analysis of Analyst's parameters gives us a basic understanding of how long it would take to pay off the initial investment of building a green building: from a low of 7.6 years at 100% occupancy rate to over 800 years if we have a 1% occupancy rate. However, this is unrealistic, becuase if we were analyze the entire data set, we find that the rate of occupancy for green to non-green buildings are significantly different. But before we continue, let us focus on the actual building operational lifecycle of 30 years.




Here we see a more accurate estimate of the payoff timetable for the building if we built it green. If we have 100% occupancy, we expect to pay it off within 7.6 years, however if we see 25% occupacy rates, we would break-even at the end of the building's estimated operational lifecycle, and any lower than 25% we would never break-even.

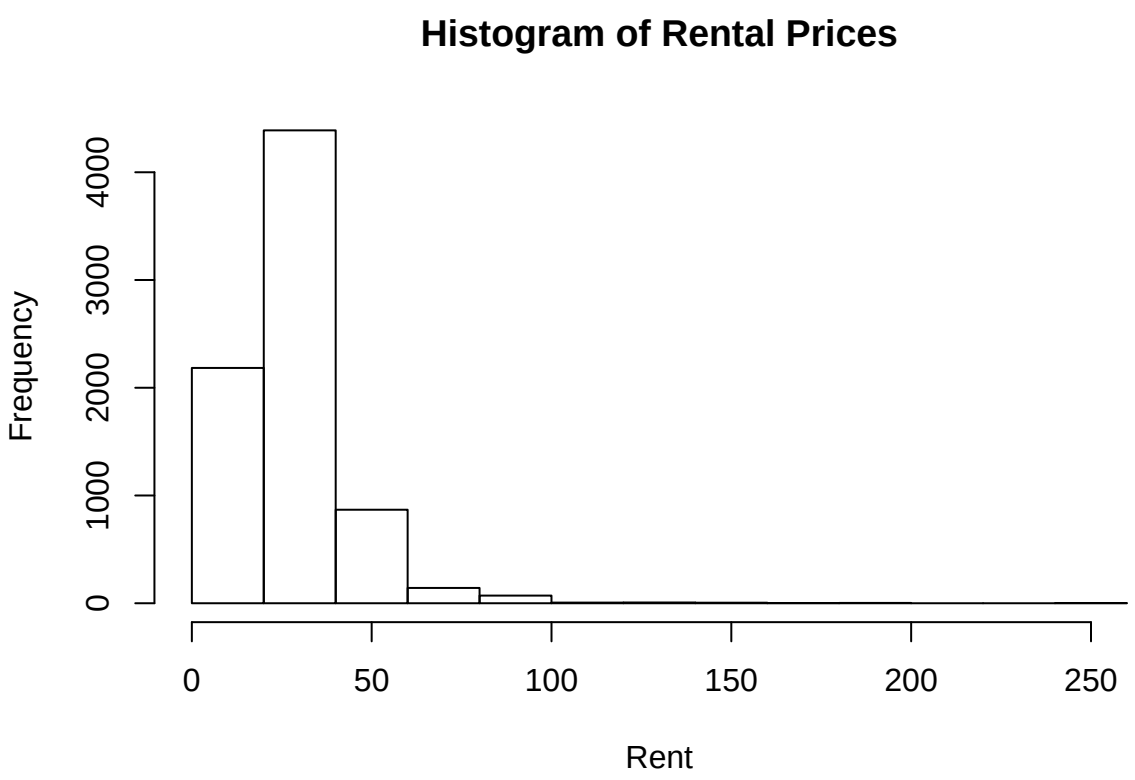
Regardless of the analysis there are a number of points of fault within the Analyst's analysis.

- One: by eliminating buildings where its occupancy rates are below 10%, we exclude important data that would give us insight as to how the market is doing, namely the fact that we don't expect occupancy rates to remain constant nor at max-capacity at the onsent of the building's commission. Becuase the data does not give us longitudial information regarding a given building's occupancy rate over the time of its lifetime, we do not have an idea of how a building going from non-green to green affects its rental value.
- Two: by utilizing the median rental value of the subsets: green and non-green, we skew the data because we have more green versus non-green buildings:

```
##
##    0    1
## 7209  685
```

Where "0" represent the number of non-green buildings and "1" represent the number of green buildings

- Three: As for the decision to utilize the median instead of the mean value of rent, a cursory assessment of different values provide different results of a green vs. non-green building.




A histogram of the Analyst's data set provides a view of how the rental prices are positively skewed. There are more buildings with rental prices in the 0-50 range than >50. Using the median value prices would have a result less than the mean value.

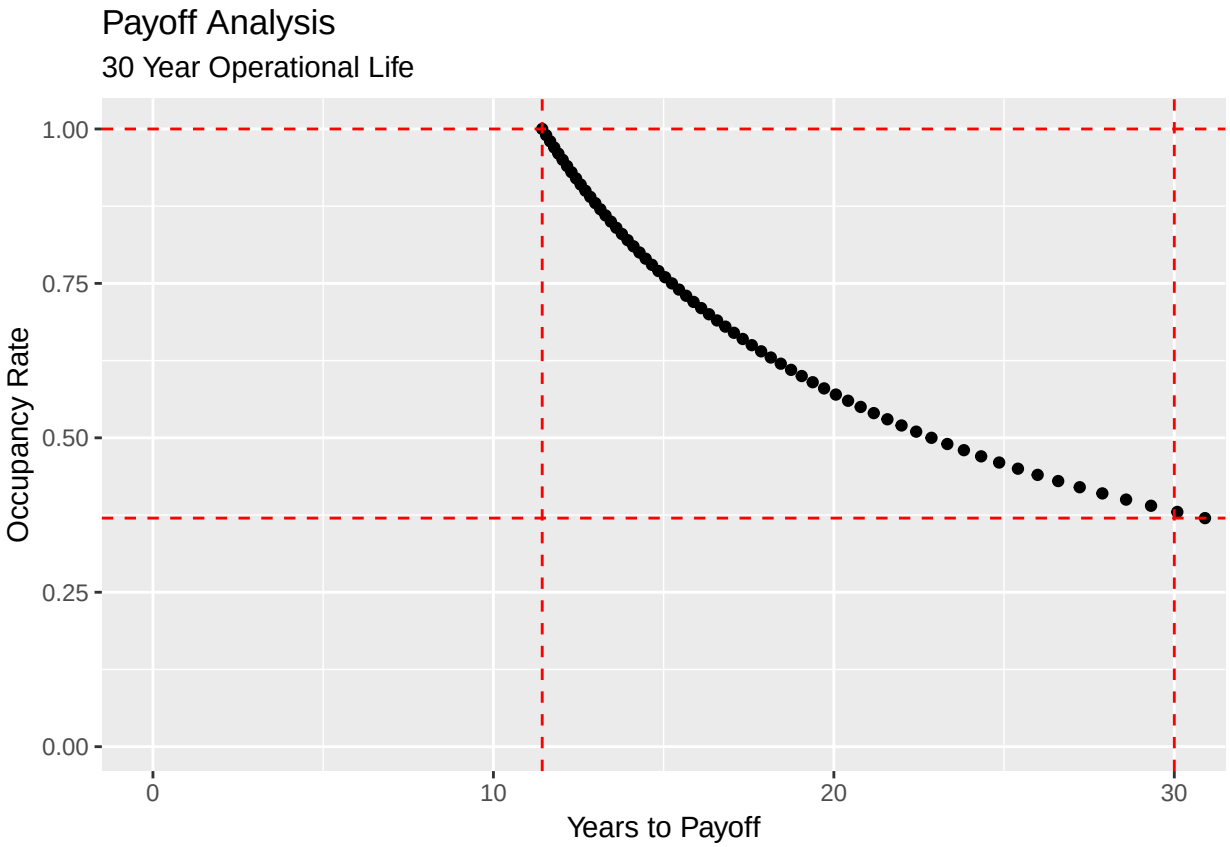
```
##                  Median     Mean
## Whole Dataset 25.29000 28.58585
```

However, if we were to compare the price differences within the green versus non-green building subsets, we see that the effect of green buildings are less than anticipated. Intead of a 2.6 difference using median-values, it will only be 1.7. This will greatly extend our anticipated break-even point.

```
##                  Median     Mean
## Whole Dataset 25.29000 28.58585
## Green         27.60000 30.01603
## Non-Green     25.00000 28.26678
```

## A Better Approach

If we were to conduct the analysis with the full set of data and the mean-value, we see a different picture.




Here we see a more accurate estimate of the payoff timetable for the building if we built it green using mean-values. If we have 100% occupancy, we expect to pay it off within 11.43 years, however if we see 37% occupacy rates, we would break-even at the end of the building's estimated operational lifecycle, and any lower than 37% we would never break-even.

In conclusion, we see that a more accurate assessment of the viability of building green is represented by using the full data set with mean-values instead of a truncated data set with median-values.

## Question 2