STA 380, Part 2: Exercises 1

Probability practice

Part A.

Here's a question a friend of mine was asked when he interviewed at Google.

Visitors to your website are asked to answer a single survey question before they get access to the content on the page. Among all of the users, there are two categories: Random Clicker (RC), and Truthful Clicker (TC). There are two possible answers to the survey: yes and no. Random clickers would click either one with equal probability. You are also giving the information that the expected fraction of random clickers is 0.3.

After a trial period, you get the following survey results: 65% said Yes and 35% said No.

What fraction of people who are truthful clickers answered yes?

Answer:

We define the following events:
RC: Random Clicker
TC:Truthful Clicker
Y: Answered yes
N:Answered no

From the problem formulation we know the following:

$P(Y) = 0.65$
$P(N) = 0.35$
$P(Y|RC) = 0.5 = P(N|RC)$
$P(RC) = 0.3$
$P(TC) = 0.7$

From the law of total probability we have that :

$P(Y) = P(Y|TC) \cdot P(TC) + P(Y|RC) \cdot P(RC)$
$0.65 = P(Y|TC) \cdot 0.7 + 0.5 \cdot 0.3$
$P(Y|TC) = \frac{0.65-0.15}{0.7}$
$P(Y|TC) = \frac{5}{7} = 0.71$

So, 71% of the people who are truthful clickers answered yes.

Part B

Imagine a medical test for a disease with the following two attributes:

The sensitivity is about 0.993. That is, if someone has the disease, there is a probability of 0.993 that they will test positive. The specificity is about 0.9999. This means that if someone doesn't have the disease, there is probability of 0.9999 that they will test negative. In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it (or 0.000025 as a decimal probability).

Suppose someone tests positive. What is the probability that they have the disease? In light of this calculation, do you envision any problems in implementing a universal testing policy for the disease?

Answer:

We define the following events:
P: Tested positive
D: Has the desease

From the problem formulation we know the following:

$P(P|D) = 0.9930$
$P(P^c|D^c) = 0.9999$
$P(P|D^c) = 1 - P(P^c|D^c) = 0.0001$
$P(D) = 0.000025$
$P(D^c) = 0.999975$

Then, in order to answer the question,we will use Bayes Rule (or conditional probability and law of total probability-it's the same):

$P(D|P) = \frac{P(DP)}{P(P)} = \frac{P(P|D)\cdot P(D)}{P(P|D)\cdot P(D)+P(P|D^c)\cdot P(D^c)} = 0.1988 \approx 0.2$

So approximately only 20% of the people who test positive, actualy have the desease.This is a very bad outcome, since a lot of people are going to get a false positive result and become worried with no actual reason. So I would suggest either implement a more accurate proceedure or suggest testing only to certain desease-high risk groups of the population.

Exploratory analysis: green buildings

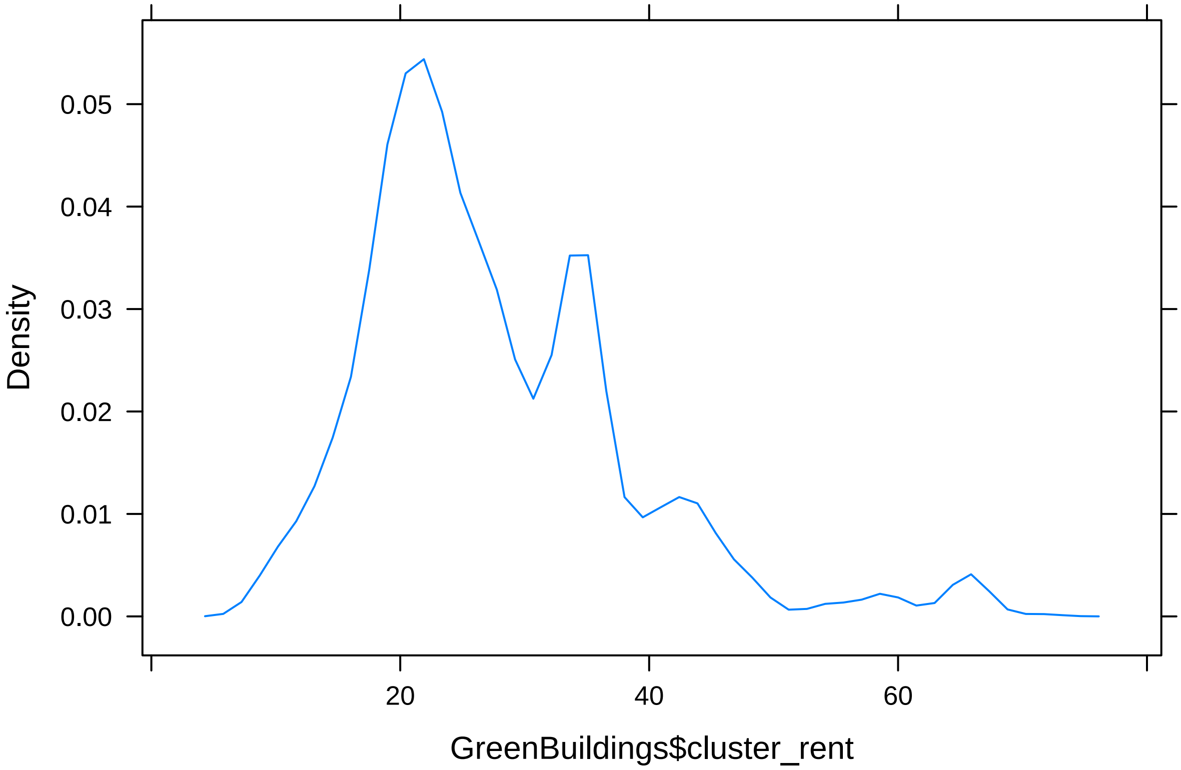
The dataset greenbuildings.csv contains data on 7,894 commercial rental properties from across the United States. Of these, 685 properties have been awarded either LEED or EnergyStar certification as a green building.In order to provide a control population for the 685 green buildings, each of these buildings was matched to a cluster of nearby commercial buildings in the CoStar database. Each small cluster contains one green-certified building, and all non-rated buildings within a quarter-mile radius of the certified building. On average, each of the 685 clusters contains roughly 12 buildings, for a total of 7,894 data points.

What we wish to achieve here is do some exploratory analysis and familiarize ourselves with the data so we can efficiently tacle the next assignment.

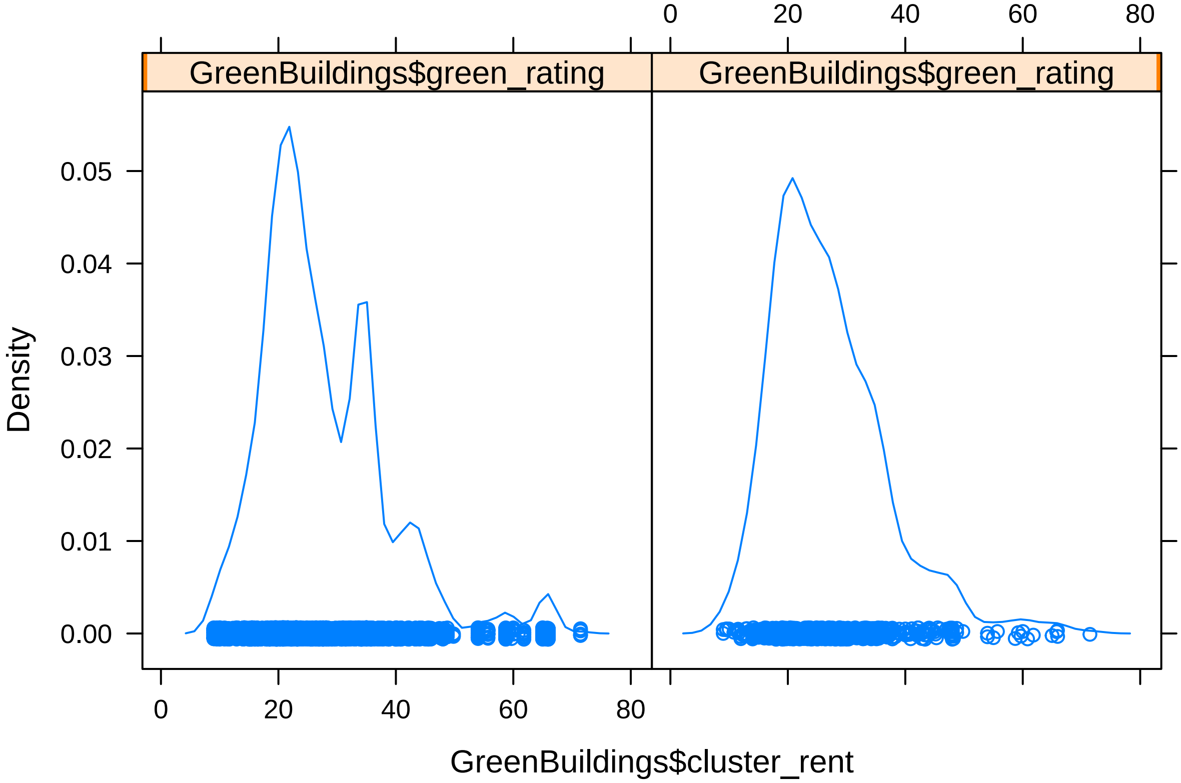
The summary of the Rent variable is:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.50   25.16   28.42   34.18  250.00
```

Here is a density plot of the cluster_rent.



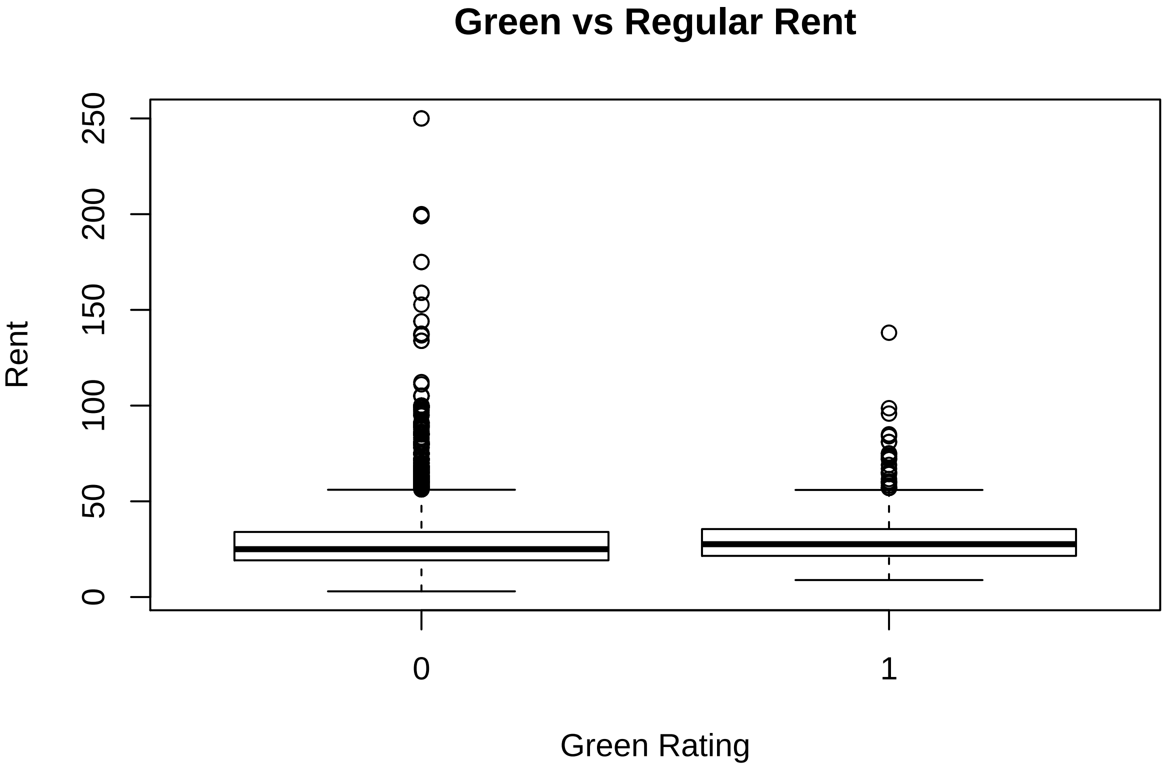
It is interesting to see the second spike and that it is created because of the non green buildings:



The low occupacy buildings that the 'Excel Guru' removed from the dataset are :

```
## [1] 23
```

Which indeed are too few, but I do not see the reason to exclude them from the analysis. If they had something 'weird' it is better to be identified and then logically excluded. For all we know it could distrort the analysis.



The assignment

The analysis presented to me by the 'Excel Guru' is not entirely mistaken,but very misleading. His analysis on whether the developer should invest in green building and go green is very myopic for the following reasons:

1) Economic perspective: The fact that after certain years the green buildings will begin to recuperate and move to produce earnings is naive. The same applies for non green constructions as well. The question should be answered either on a predifined horizon (10,20 years where we would be able to compare the earnings of each investment separately), or using some notion of opportunity cost; for example if the developer was to invest these 5 million somewhere else instead of getting a green certification,would be benefitted financially or not? Also we would have to take into account the maintainance of the 2 building types.
   Having said that, is clear that a green building saves the user a lot of money in the long run, but the effect it has on the landlord/developer needs further evaluation.

2) The way conclusions were drawn from the data analysis. I will analyze the most important in my opinion. The expert,disregarded entirely has possibility of confounding variables for the relationship between rent and green status and did all his analyses separately for green and non green buildings.

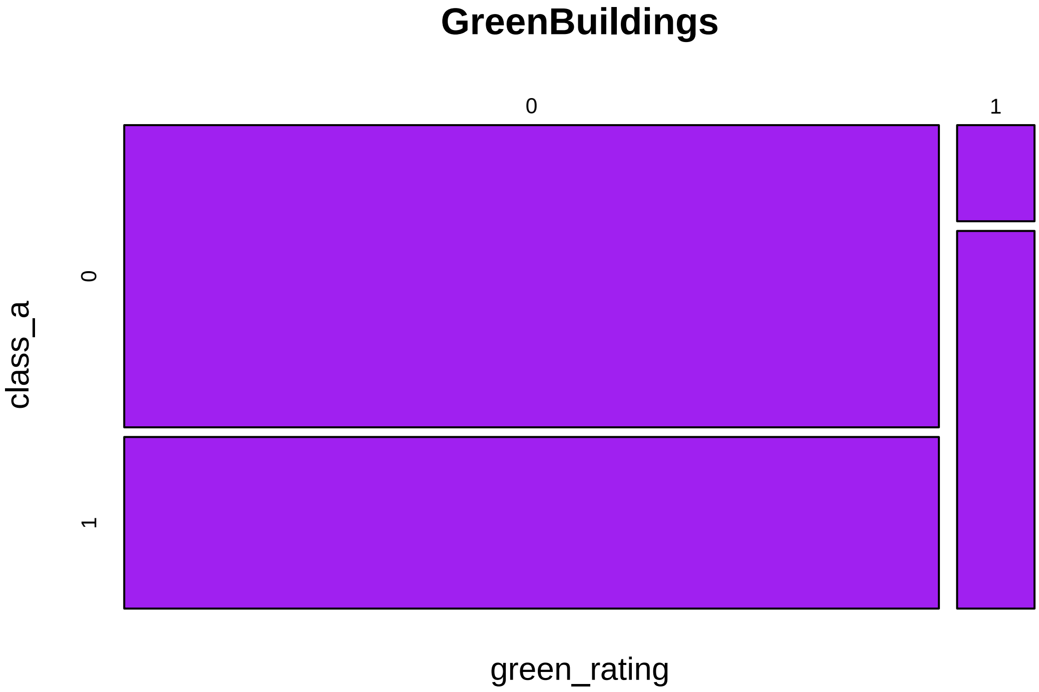
But as we can see, one such is the class (A,B) distribution among the buildings.

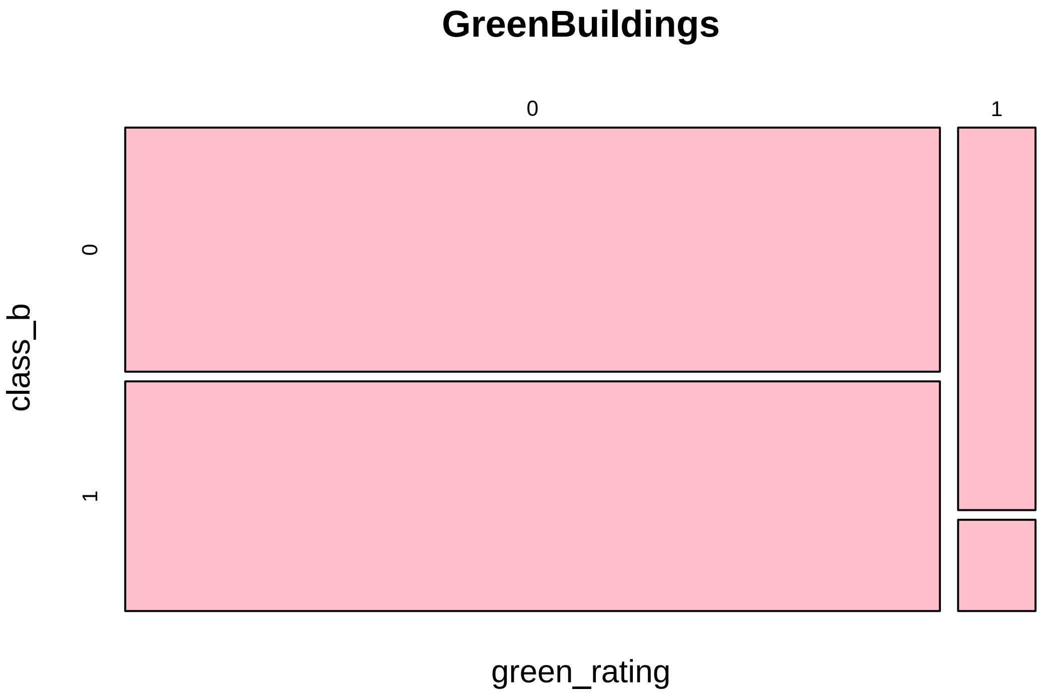
```
##                       class_b    0    1
## green_rating class_a
## 0            0               1103 3495
##              1               2611    0
## 1            0                  7  132
##              1                546    0
```

So we that approximately 80% of the green buildings is classified as class A and 19% is classified as Class B, whereas for the rest of the buildings only 33% and 44% is class A buildings and class B respectively.

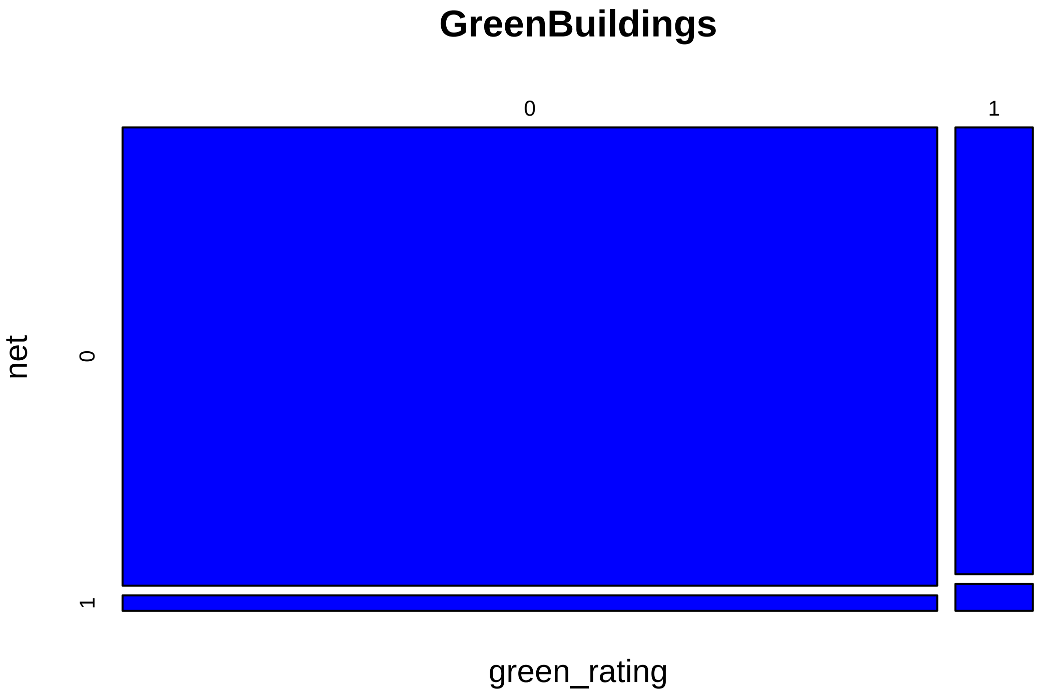
After that it is clear that class is a confounding variable between rent and green buildings.

If we wish to display the allocation of Class A and class B buildings we can see the next two plots.





Another confounding factor is the net variable:



The rent alone is not a good inditator of future profit. The vast majority of greenbuildings have a net contract and their utilities are included in the rent. So the actual rent is expected to be lower. This is something that the developer would like to take into account.

To sum up, I think that a more carefull analysis needs to be done in order to say whether or not the developer should proceed into investing into the making of a green building. This research is beyond the scope of this exercise. I strongly dissagree with the kind of analysis the expert decided to conduct both from a statistical point as well as economical. What I would stress as a main mistake is that the existence of confounding variables for the relationship between rent and green status were overlooked and this affects the study. It could be that the green buildings are naturally pricier since they are newer,bigger,classier (Class A,B) and the utilities are included in the rent.

Bootstrapping

We need to consider the following five asset classes, together with the ticker symbol for an exchange-traded fund that represents each class:
SPY,TLT,LQD,EEM,VNQ. Utilising real time data on the above ETF's we will try to estimate the risk/return of these assets. The final step is to consider having $100,000 to invest in one portfolio. These portfolios are an even split between the 5 assets, a safer option than the split and a more aggressive. The main goals here are:

1. To quantify the risk/return properties of the five major asset classes listed above.

2. Explain the choices of the "safe" and "aggressive" portfolios.

3. Use bootstrap resampling to estimate the 4-week (20 trading day) value at risk of each of the three portfolios at the 5% level.

Portfolio 1: The even split

We import 10 year's worth of data from the stock market, to use in our analysis.

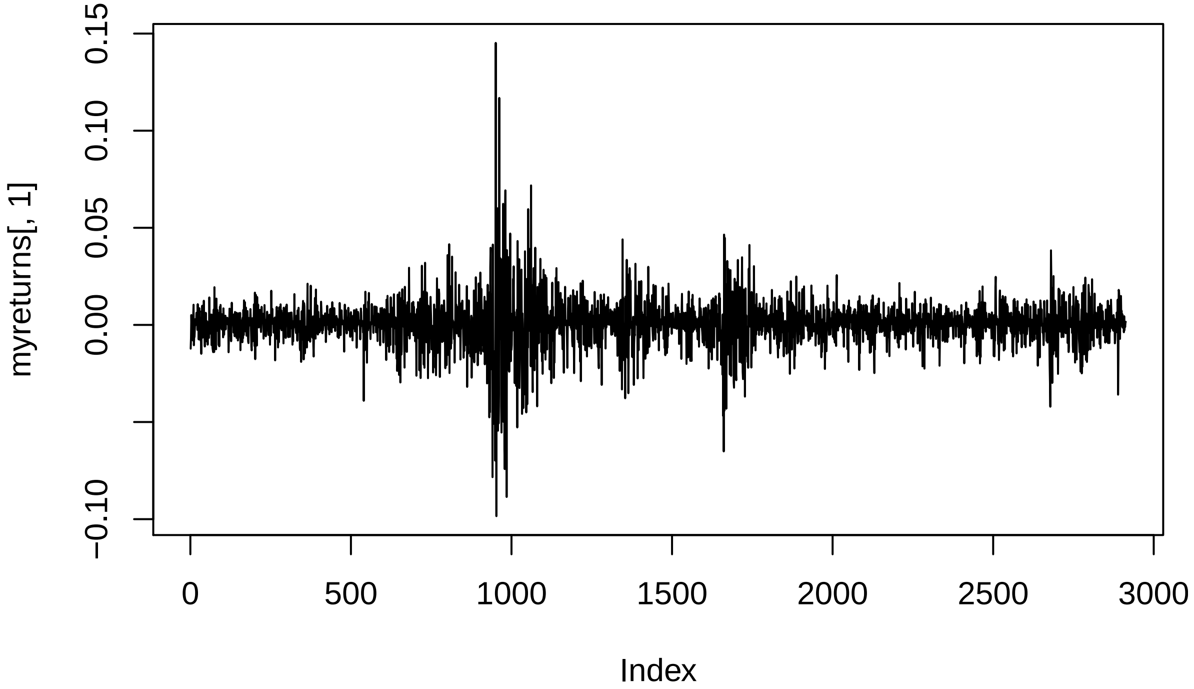
The first few rows of our dataset can be seen here:

```
## GMT
##            SPY.Open SPY.High SPY.Low SPY.Close SPY.Volume SPY.Adj.Close
## 2005-01-03   121.56   121.76  119.90    120.30   55748000      95.29625
## 2005-01-04   120.46   120.54  118.44    118.83   69167600      94.13178
## 2005-01-05   118.74   119.25  118.00    118.01   65667300      93.48222
## 2005-01-06   118.44   119.15  118.26    118.61   47814700      93.95751
## 2005-01-07   118.97   119.23  118.13    118.44   55847700      93.82284
## 2005-01-10   118.34   119.46  118.34    119.00   56563300      94.26645
##            TLT.Open TLT.High TLT.Low TLT.Close TLT.Volume TLT.Adj.Close
## 2005-01-03    88.18    88.84   88.16     88.74    1168000      58.23352
## 2005-01-04    88.72    88.75   87.81     87.81    1935400      57.62323
## 2005-01-05    87.99    88.55   87.94     88.28    1094100      57.93165
## 2005-01-06    88.29    88.54   88.22     88.34    1057400      57.97103
## 2005-01-07    88.76    88.87   88.35     88.54     738700      58.10227
## 2005-01-10    88.64    88.72   88.47     88.68     379400      58.19415
##            LQD.Open LQD.High LQD.Low LQD.Close LQD.Volume LQD.Adj.Close
## 2005-01-03   111.71   112.25  111.50    112.25    1497800      66.89484
## 2005-01-04   112.27   112.29  111.50    111.62      90200      66.51940
## 2005-01-05   111.65   111.91  111.46    111.71     120700      66.57303
## 2005-01-06   111.55   111.95  111.55    111.79      43900      66.62071
## 2005-01-07   111.91   111.92  111.43    111.74      68300      66.59091
## 2005-01-10   111.76   111.76  111.40    111.55      73200      66.47768
##            EEM.Open EEM.High EEM.Low EEM.Close EEM.Volume EEM.Adj.Close
## 2005-01-03   201.70   202.45  199.38    199.75    4275000      18.03730
## 2005-01-04   199.25   199.35  193.60    193.60    4205700      17.48196
## 2005-01-05   193.40   193.77  191.20    191.23    3006900      17.26795
## 2005-01-06   191.85   192.12  190.13    191.10    2268000      17.25621
## 2005-01-07   192.40   192.76  190.50    191.47    4920300      17.28962
## 2005-01-10   192.60   193.61  191.65    191.71    2007000      17.31129
##            VNQ.Open VNQ.High VNQ.Low VNQ.Close VNQ.Volume VNQ.Adj.Close
## 2005-01-03    56.75    56.75   55.50     55.89      31900      33.22696
## 2005-01-04    55.89    56.28   55.05     55.05      52500      32.72758
## 2005-01-05    55.17    55.17   52.83     53.22      77300      31.63963
## 2005-01-06    53.15    53.82   53.15     53.63      42200      31.88338
## 2005-01-07    53.87    53.88   53.27     53.51      24200      31.81204
## 2005-01-10    53.45    53.65   53.20     53.34      12500      31.71097
```

Also here we can see the calculated returns from the closing prices of the assets.

```
##            SPY.PctReturn TLT.PctReturn LQD.PctReturn EEM.PctReturn
## 2005-01-04  -0.012219463 -0.0104800469 -0.0056124508 -0.0307883143
## 2005-01-05  -0.006900613  0.0053524770  0.0008062761 -0.0122416480
## 2005-01-06   0.005084304  0.0006796284  0.0007161459 -0.0006800459
## 2005-01-07  -0.001433254  0.0022640275 -0.0004472934  0.0019361727
## 2005-01-10   0.004728113  0.0015811945 -0.0017003223  0.0012535267
## 2005-01-11  -0.006890755  0.0058637513  0.0023307372 -0.0018777916
##            VNQ.PctReturn
## 2005-01-04  -0.015029512
## 2005-01-05  -0.033242486
## 2005-01-06   0.007703883
## 2005-01-07  -0.002237592
## 2005-01-10  -0.003176942
## 2005-01-11  -0.010123752
```

When there is a negative sign it means we lost the indicated amount of money . Below you can see a plot of returns over time. You can see the spikes/the periods where there was a lot of volatility in the market.



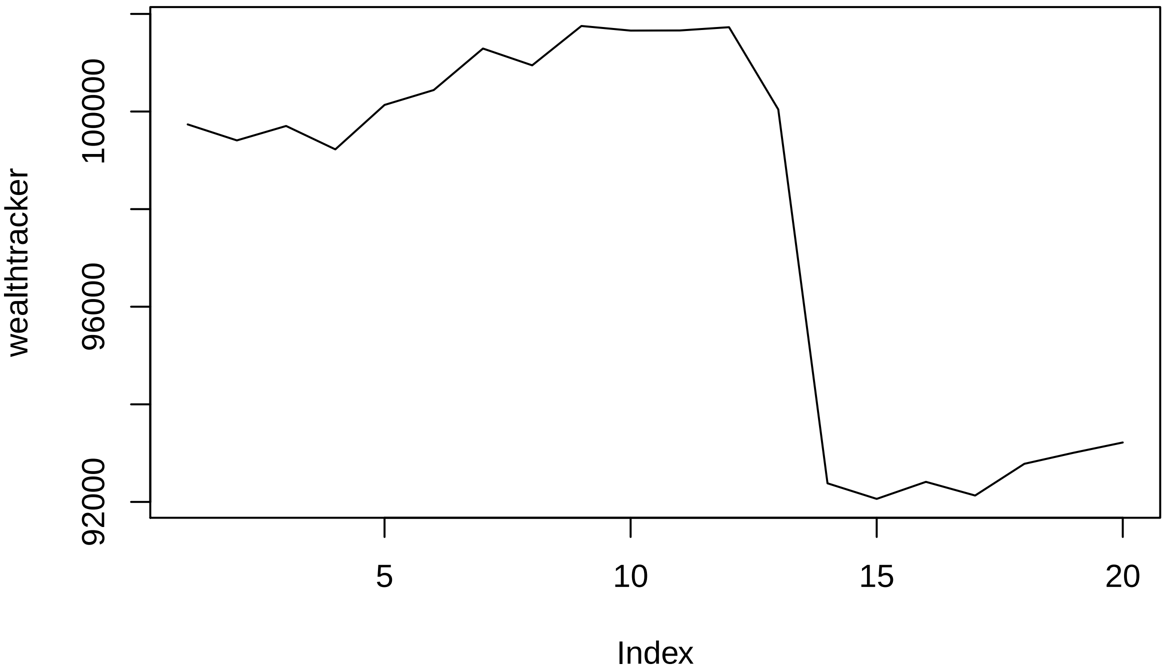
Let's assume we distribute evenly our wealth to these 5 assets. Then my return today would be:

And my new wealth would be:

```
## [1] 99625.7
```

So I earned 74 cents! I we do this for each of the 20 days our 4 week period has, then we will end with a new total wealth of

```
## [1] 93218.17
```



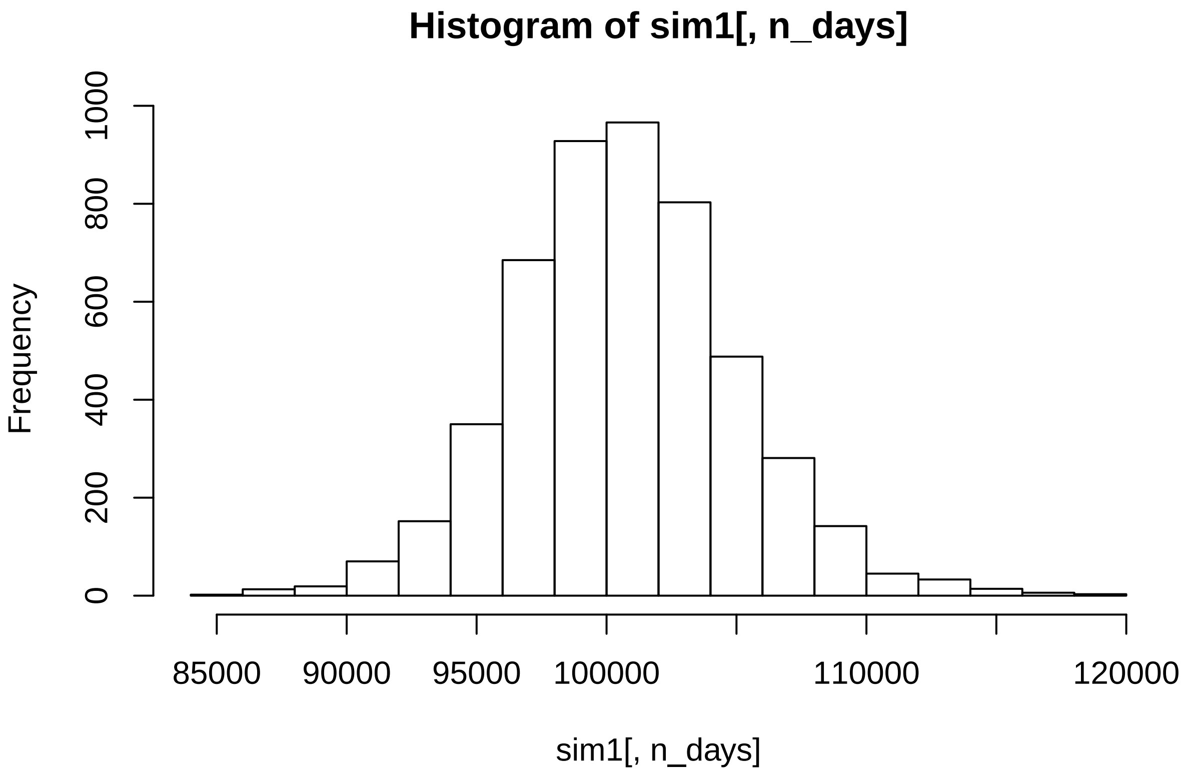
Also we can see the plot which illustrates our wealth for these past 20 days.

Now if we run a Monte Carlo simulation the summary of our results is a mean which can be interpreted as our gain in dollars from our investment and the second one is a standard deviation. We can also observe the histogram of our simulations, over the days.

```
#Result summary
head(sim1)
```

```
##               [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## result.1  98730.70  99086.98  99407.89  98311.85  98061.12  98333.83
## result.2  99765.85  99703.71  99437.10  99393.79  99654.66  99432.20
## result.3 100128.82  99673.23  99806.83 100052.71 100873.12 101909.20
## result.4  99970.72 100410.85 100850.88 100673.68 100430.86 100049.30
## result.5  99679.01  99497.12  99291.22  99337.92  99450.44  99960.72
## result.6  99997.42  98812.21  98736.95  98819.92  98818.32  98913.15
##               [,7]      [,8]      [,9]     [,10]     [,11]     [,12]
## result.1  98528.12  98661.11  98744.29  98969.41  98810.11  98445.46
## result.2  99540.22  99569.54  98773.52  98559.15  98566.44  97045.61
## result.3 101806.84 101779.96 102114.84 102056.39 101869.53 101626.31
## result.4 100232.51 100506.42 100849.35 100591.42 102921.87 104224.28
## result.5 100205.91 100602.72 101040.79 100372.74  98180.90  98250.59
## result.6  98558.46  98462.43  99201.61  99996.54  99176.19  97232.42
##              [,13]     [,14]     [,15]     [,16]     [,17]     [,18]
## result.1 103682.43 103403.79 104088.13 105099.95 103469.72 103576.54
## result.2  98122.58  97514.70  97854.07  98162.79  99068.06  99097.67
## result.3 101008.32 100635.80 100193.56  99738.01  98924.03 100028.08
## result.4 102906.80 103914.60 104144.62 104174.01 102485.80 102730.22
## result.5  98258.73  98346.38  97568.33  97393.17  97935.72  97692.79
## result.6  97518.99  97920.07  98162.65  98733.60  99385.38  99639.58
##               [,19]     [,20]
## result.1 104668.87 105075.44
## result.2  98762.31  96768.75
## result.3 100039.44  99837.76
## result.4 101381.40  96112.08
## result.5  98224.21  97362.51
## result.6  99291.50  99255.89
```

```
hist(sim1[,n_days],25)
```

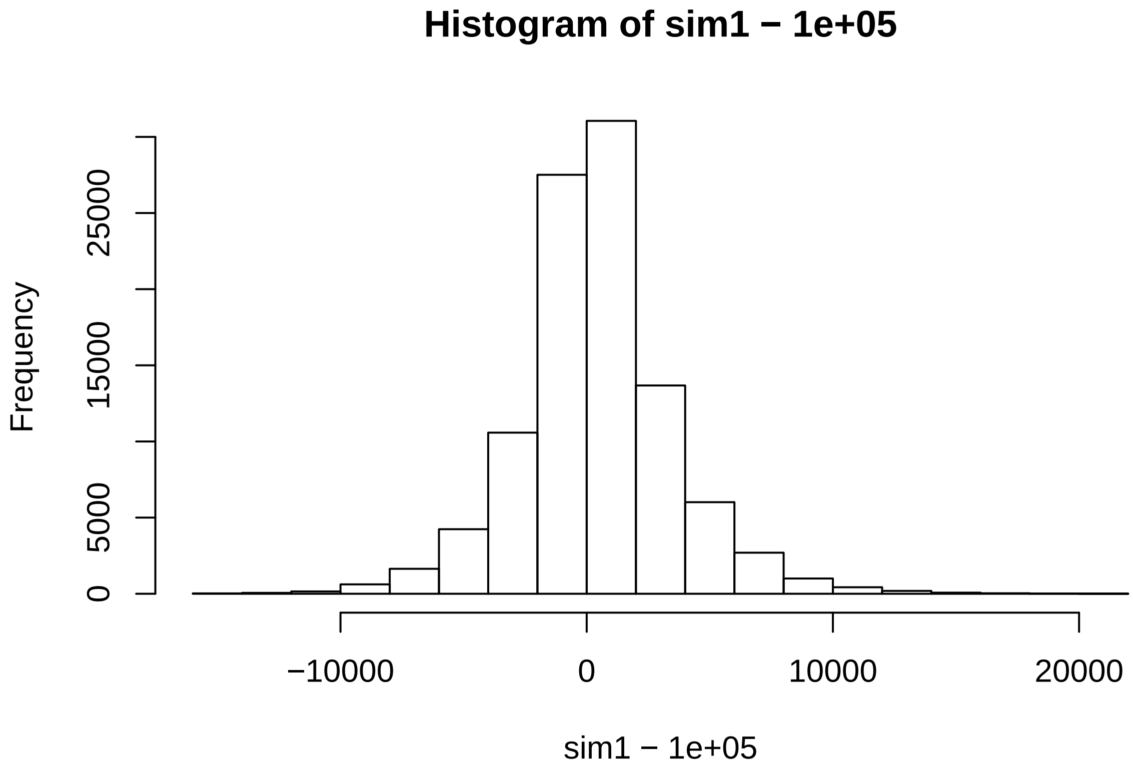


```
mean(sim1)
```

```
## [1] 100386.9
```

```
sd(sim1)
```

```
## [1] 3123.881
```



Next we calculate the 5% Value At Risk (VAR), which is the 5% quantile of the profit/loss distribution of a portfolio for 20 days.

In our case, the VAR is:

```
##        5% 
## -6054.618
```

```
##         p  quantile
##     0.050 -4632.119
```

We highlight in the graph below with blue, the VAR of our portfolio and for comparison purposes the expected return of our portfolio in this case is marked with purple. (for some reason,not knitting,while it is running,so I will not include this graph with the coloured lines)

Portfolio 2: The safe option

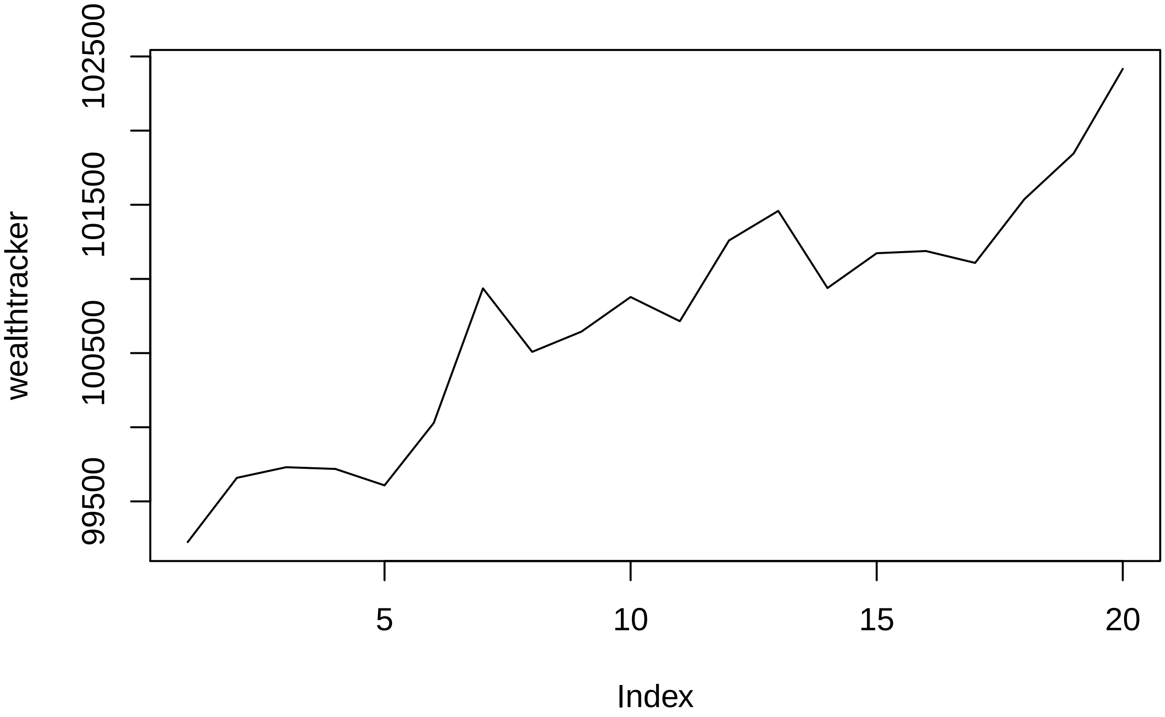
If I distribute my wealth : $40\%, 20\%, 40\%, 0\%, 0\%$

And my new wealth would be:

```
## [1] 100312.6
```

So I earned or lost the below amount of dollars! I we do this for each of the 20 days our 4 week period has, then we will end with a new total wealth of

```
## [1] 102416.2
```



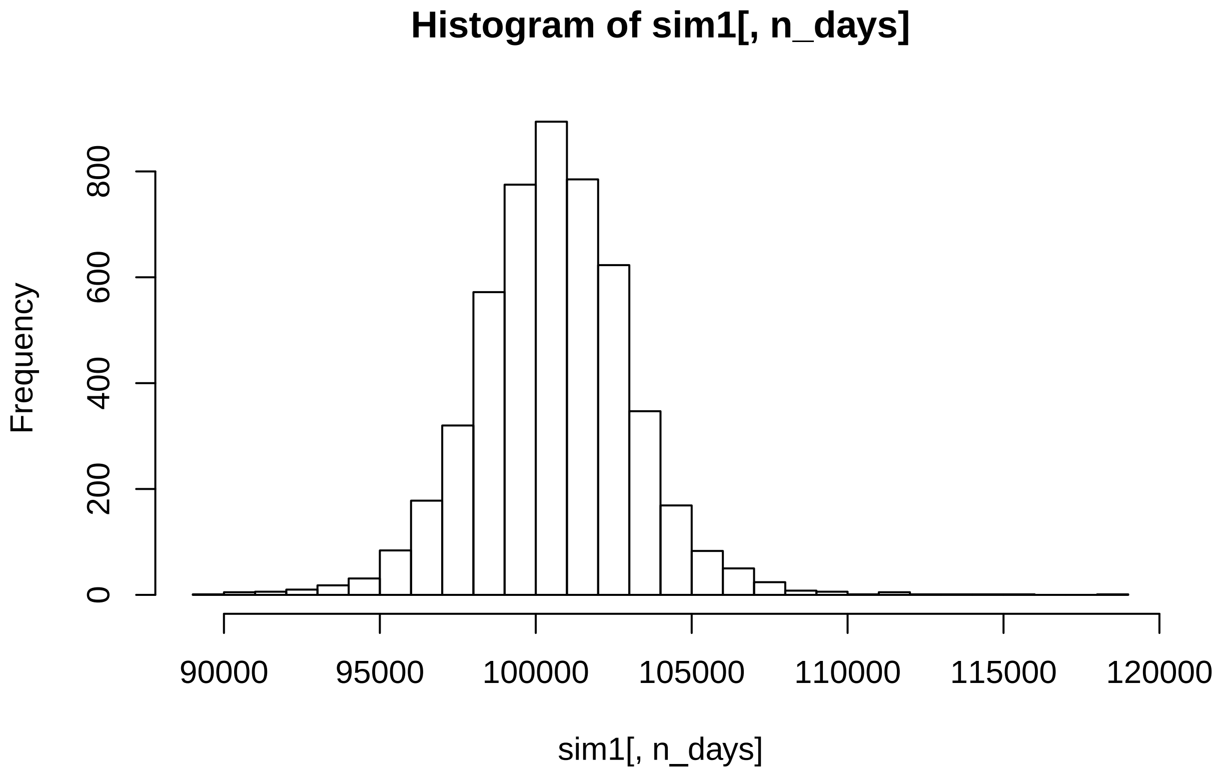
Also we can see the plot which illustrates our wealth for these past 20 days.

Now if we run a Monte Carlo simulation the summary of our results is a mean which can be interpreted as our gain in dollars from our investment and the second one is a standard deviation. We can also observe the histogram of our simulations, over the days.

```
#Result summary
head(sim1)
```

```
##              [,1]     [,2]      [,3]      [,4]      [,5]      [,6]
## result.1 100199.0 100570.8 100861.54 101183.02 101594.47 100887.59
## result.2 100582.2 100593.4 100872.83 100771.25 100685.96 100544.29
## result.3 101086.3  98906.4  98987.84  98623.72  98837.79  99375.11
## result.4 100493.8 100603.2 101040.10 100937.28 100318.39 100094.64
## result.5 100408.0 100446.7 100700.45 100796.71 100546.72 100553.17
## result.6 100353.6 100027.8 100752.25  99473.76  99514.62  99653.02
##               [,7]      [,8]      [,9]     [,10]     [,11]     [,12]
## result.1 100803.98 101468.60 101240.68 101071.92 100513.93 100781.05
## result.2 100772.15 101021.25 101519.10 101953.53 101771.85 101276.08
## result.3  98952.29  98972.74  98725.76  98332.74  97961.27  97907.51
## result.4  99854.21 100309.63 100335.07 100068.28  99379.88  99139.72
## result.5 101054.15 100952.16 101032.90 101177.84 101529.23 101285.36
## result.6  99349.07  99451.35  99732.42  99506.08  99218.51  99163.37
##              [,13]     [,14]     [,15]     [,16]     [,17]     [,18]
## result.1 100882.29 101078.69 101172.30 100998.95 100958.05 100957.91
## result.2 101335.64 100829.96 100892.54 100751.60 100538.46 100176.23
## result.3  97324.46  96703.37  96676.30  96733.41  97186.41  97255.09
## result.4  99043.24  99126.53  99428.11  99478.83 100381.51 101047.59
## result.5 100980.87 100774.24 100952.80 101141.48 101306.40 101552.85
## result.6  97827.89  98276.99  98519.29  99624.46  99961.81  99868.78
##              [,19]     [,20]
## result.1 101964.22 102115.88
## result.2 100842.03 101551.71
## result.3  97359.01  97729.43
## result.4 101191.03 101149.17
## result.5 101139.73 100813.04
## result.6 100114.49  99220.36
```

```
hist(sim1[,n_days],25)
```

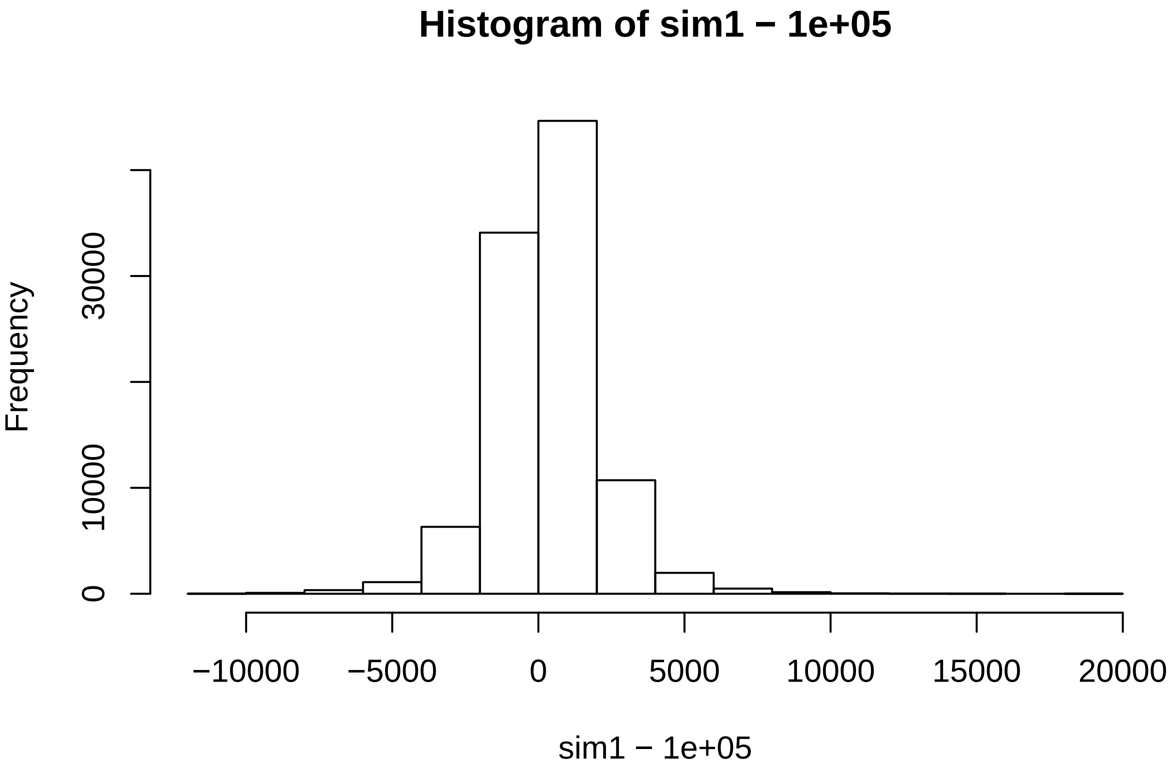


```
mean(sim1)
```

```
## [1] 100293.2
```

```
sd(sim1)
```

```
## [1] 1822.405
```



Next we calculate the 5% Value At Risk (VAR), which is the 5% quantile of the profit/loss distribution of a portfolio for 20 days.

In our case, the VAR is:

```
##        5% 
## -3389.938
```

```
##         p  quantile
##     0.050 -2526.853
```

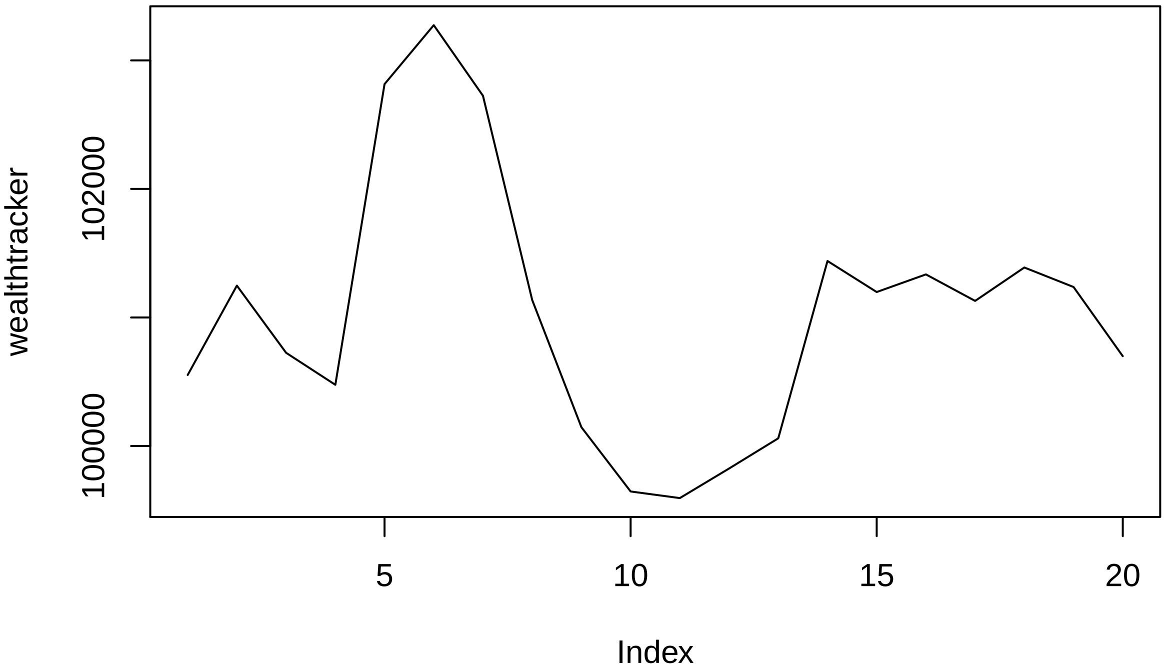
Or we can use again and again the first model by re-allocating our investments: For example: If I distribute my wealth : $20\%, 10\%, 10\%, 20\%, 40\%$

And my new wealth would be:

```
## [1] 99879.38
```

So I earned 74 cents! I we do this for each of the 20 days our 4 week period has, then we will end with a new total wealth of

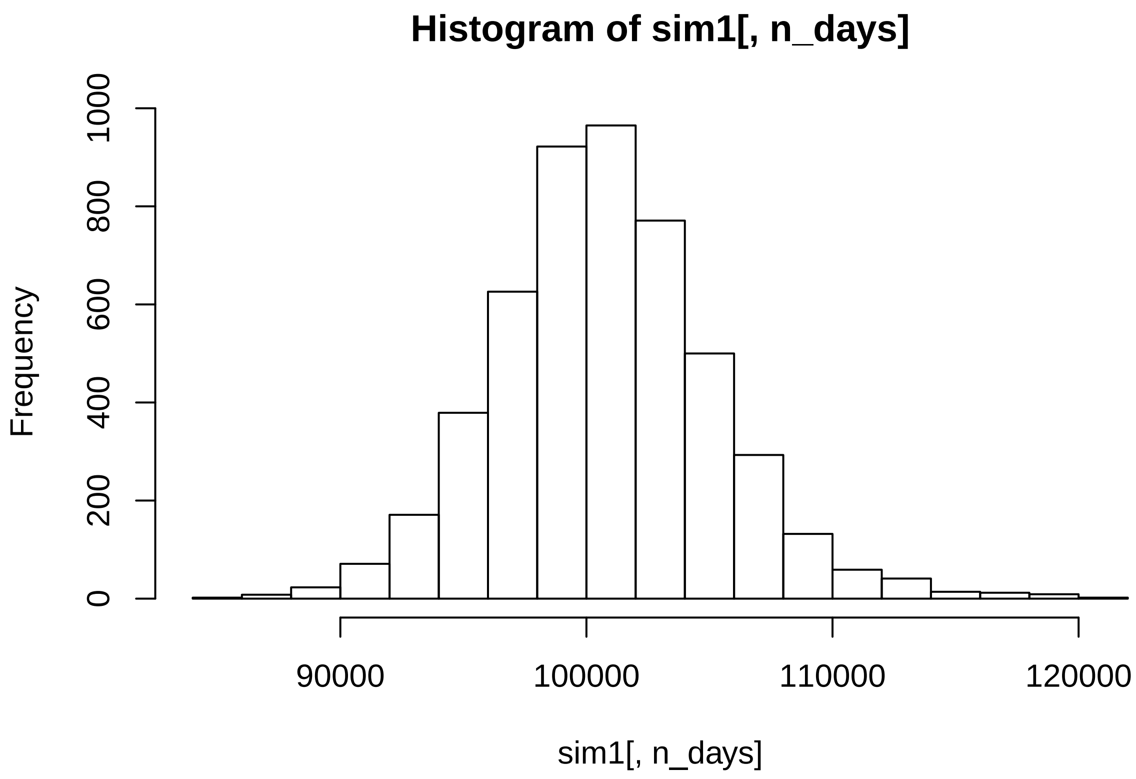
```
## [1] 100698.8
```



Also we can see the plot which illustrates our wealth for these past 20 days.

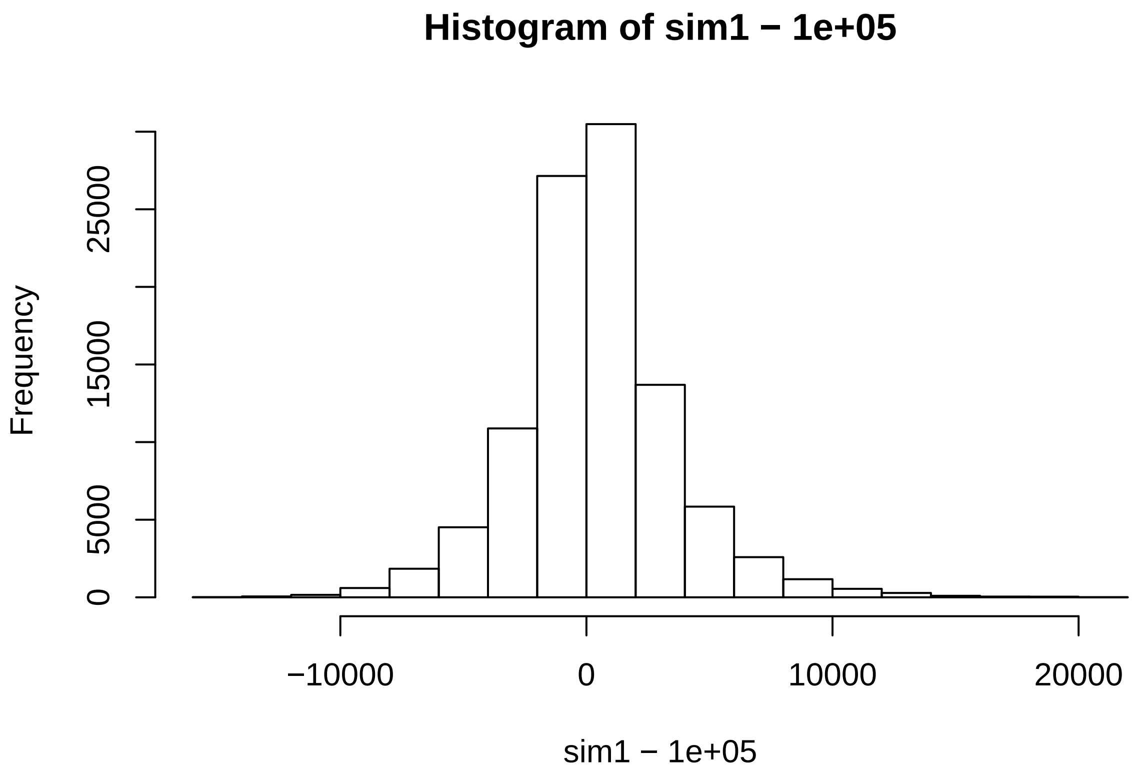
Now if we run a Monte Carlo simulation the summary of our results is a mean which can be interpreted as our gain in dollars from our investment and the second one is a standard deviation. We can also observe the histogram of our simulations, over the days.

```
##               [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## result.1  99763.26  97012.02  96554.78 100211.78 100062.25  99444.25
## result.2  99620.35  99997.17  99527.29  99401.21  99855.49 100100.62
## result.3  98094.29  98441.56  98126.41  98348.71  99113.81  97346.58
## result.4 100084.75 100254.10  99484.59  99245.00  99801.73 100893.09
## result.5 100497.11 100985.66 100456.78 100070.11 102097.51 101580.26
## result.6  99870.67 100605.00 100742.57 101043.68 100988.71 101118.75
##               [,7]      [,8]      [,9]     [,10]     [,11]     [,12]
## result.1  99802.93 100223.62 100078.21  98182.19  98044.58  97993.20
## result.2 103071.72 102937.66 106084.47 105439.76 104445.64 104556.85
## result.3  97347.93  98042.94  97979.59  97606.47  95784.97  91834.61
## result.4 101375.63 100990.81 101507.25 101765.94 101428.25 100858.05
## result.5 101755.00 102060.53 102823.97 102514.29 102635.96 102666.22
## result.6 101386.13 101659.49 101459.15 100345.53  99630.22 101436.41
##              [,13]     [,14]     [,15]     [,16]     [,17]     [,18]
## result.1  97998.54  98784.96  99150.26  98778.23  99940.52  99747.58
## result.2 105134.01 105640.25 109020.17 108388.89 108742.76 108966.72
## result.3  91110.60  89724.03  89880.58  89571.57  90956.91  90857.20
## result.4 100830.50 100881.23 101498.00 101626.27 100694.73 100903.37
## result.5 102219.45 102102.39 102168.49 102222.18 101791.51 101025.23
## result.6 101141.21 101806.34 101282.12 100336.22 100350.79 100985.73
##              [,19]     [,20]
## result.1  99472.98  99157.63
## result.2 108128.32 108160.47
## result.3  90555.92  90103.49
## result.4 100000.35  99820.79
## result.5 101269.58 101580.67
## result.6 102371.27 102594.94
```



```
## [1] 100382.3
```

```
## [1] 3226.891
```



Next we

calculate the 5% Value At Risk (VAR), which is the 5% quantile of the profit/loss distribution of a portfolio for 20 days.

In our case, the VAR is:

```
##        5%
## -6278.173
```

```
##        p quantile
##     0.05 -4749.02
```

Concluding, the assets are sorted from safest to riskiest, and whether the investor will chose a certain portfolio, this depends on whether he/she prefers to take risks or not, the horizon available etc. The portfolio 1 is ideal if someone wants to invest for a long period of time and then cash out. It is very safe,but the rewards are very slow. The portfolio 3 is the opposite, where the reward is high but the risk of losing money is high as well. In my opinion one should not put all of his/her money only in one bucket. I would go with Portfolio 2 which is a balanced approach of the 2 worlds.

Market segmentation

To TA:
I began trying a PCA approach for this exercise. My main problem was that I could not interprete the results. Now I am trying to use kmeans++ for this exersise. So far I am using small clusters because they are more interpretable.