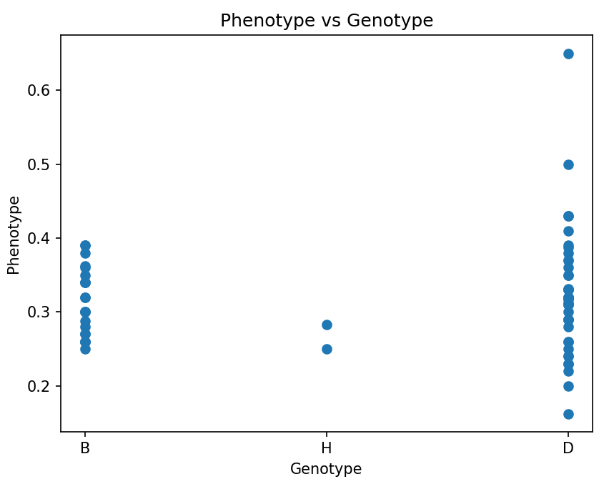
# Systems Genetics – Ex2

* Checking phenotypes.xls using pandas we can add a column of empty values count in a row and the standard deviation (std) of the rows values.  
  Then we sort for high std and low empty count that contain Pubmed Id.  
  (See find\_relevant\_phenotype func in code file).One of the top values we get is:  
  **Morphine response (50 mg/kg ip), locomotion from 0-180 min (total activity over 3 hour test) after injection in an activity chamber for females [cm]**  
  ID\_FOR\_CHECK: 1231  
  Pubmed ID: 19958391
* Confusingly I thought I should use a SNP with the same ID\_FOR\_CHECK and saw there are no breeds with H and a valid phenotype value (not Nan). So I checked for different phenotypes by changing the function (find\_relevant\_phenotype) to consider the genotype as well to find phenotypes with low empty values count that also have H in the genotypes (now ignoring std), we get:  
  **Pain sensitivity, vocalization threshold to mild foot shock for females [mA]**  
  ID\_FOR\_CHECK: 1195  
  Pubmed ID: 19958391

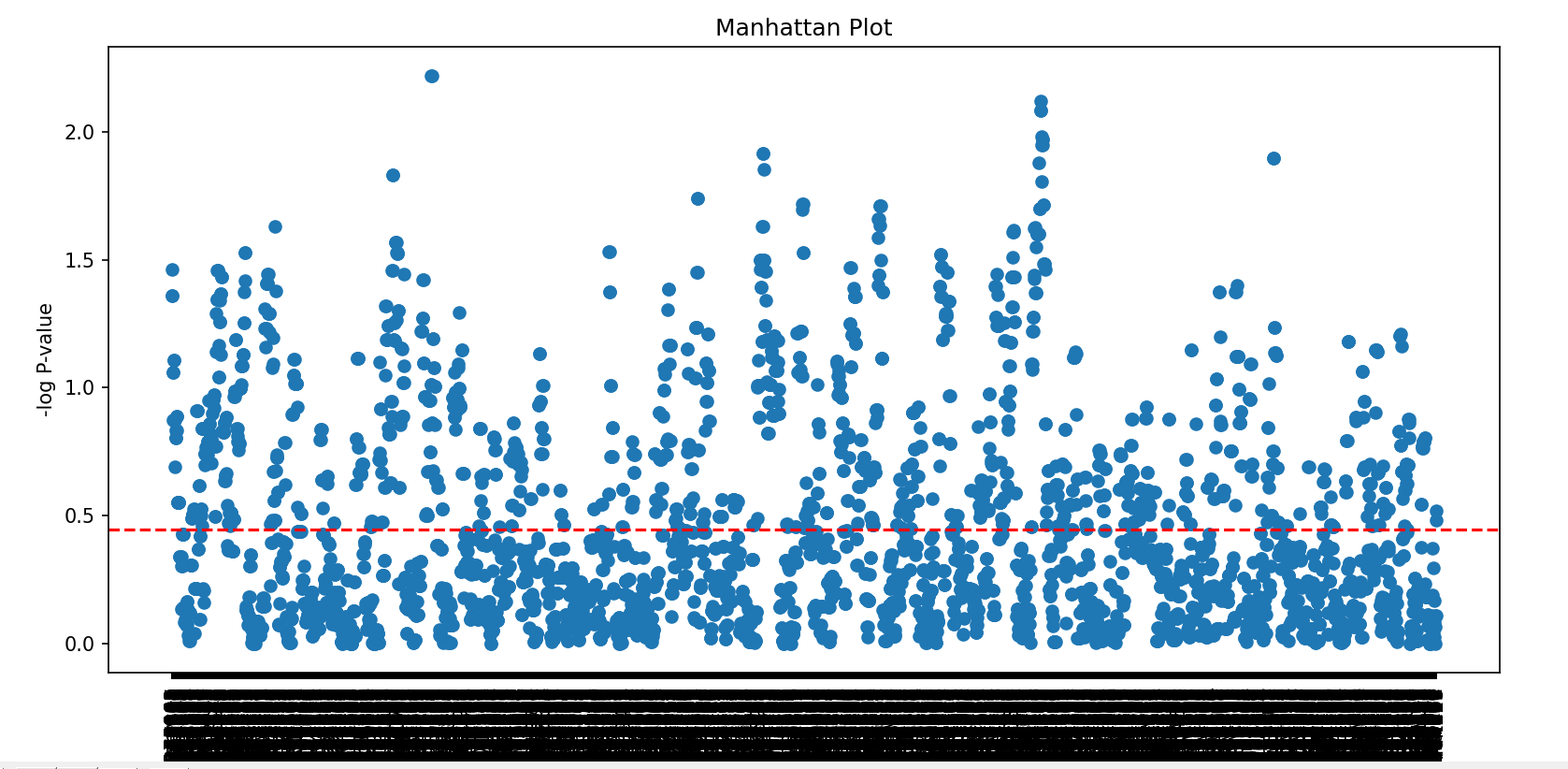
1. A picture containing text, screenshot, display, number

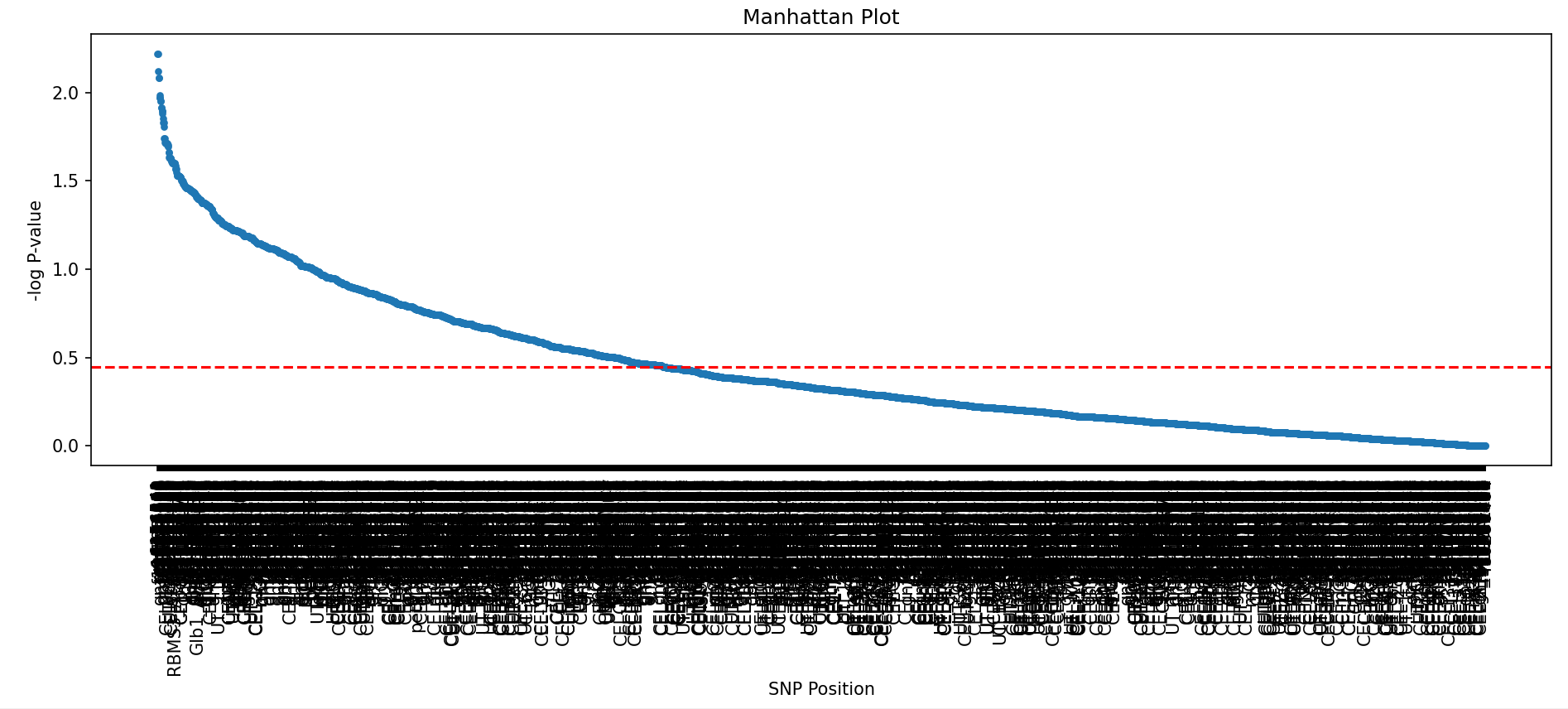
   Description automatically generatedDue to filtering beforehand, ID\_FOR\_CHECK = 1195 will be used as our phenotype and our SNP for the analysis (It has heterozygotes). SNP name: gnf05.066.286  
   Now we can check the requested models using what we learnt and python:  
   First plotting the data we see it’s not so good for linear regression:  
     
     
     
     
     
     
     
     
     
     
     
      
     
     
   1. We consider cases with only B or D genotype.  
      Given linear regression as we learnt:

In addition we get that our calculation match the functions from python’s stats models 😊

* 1. This time we consider cases with only B or D or H genotype.  
     Given linear regression as we learnt:
  2. This time we consider only cases with B or D genotype and we are interested to do ANOVA test.   
     A picture containing text, font, screenshot

     Description automatically generatedIn addition it matches the results of the python stats model ANOVA func 😊
* Comparing the results we got in all tests we can see that (a) and (c) are equivalent and it makes sense by the theory because ANOVA is a generalization to t-test with several groups and here we have only 2 groups so it’s similar to regular t-test which we have in the linear regression situation in (a).  
  Regarding (b) we can see that adding H to consideration gives almost similar regression params beta1 and beta0. The model is slightly improved but still the p-value is insignificant and the R squared value is very low so the model is not good in this case either.

1. The code can be found in the py file, the relevant funcs are:  
   q\_2\_analysis, prep\_data, regression\_model and plot\_q2\_results.  
   (There are comments above the funcs: related to Q2)  
     
     
   Manhattan plot (the red line is the mean of -log(p-value)):  
     
     
     
     
     
     
     
     
     
     
   Another view after sorting the values:



* The best-scored SNP is rs6156541 with a -log(p-value) of 2.22.  
  If we convert it back to p-value we get 0.006.  
  If we consider p-value lower than 0.05 and use Bonferroni correction (dividing 0.05 by number of SNPs we have 3796) we get that we need p-value lower than or in terms of . So we get that none of the SNPs is significant to the chosen phenotype.
* We can see that the SNP we used for Q1 got here:  
   gnf05.066.286: 0.19288597970434676  
  Which is p-value: 0.641 🡪the value we got in Q1 😊
* Unfortunately, in the relevant pubmed:  
  https://pubmed.ncbi.nlm.nih.gov/19958391/  
  There is no discussion about the specific phenotype I chose, it is just mentioned as one parameter among many others that was measured. It is possible that this phenotype didn’t get significant results based on the genotype and thus wasn’t discussed.