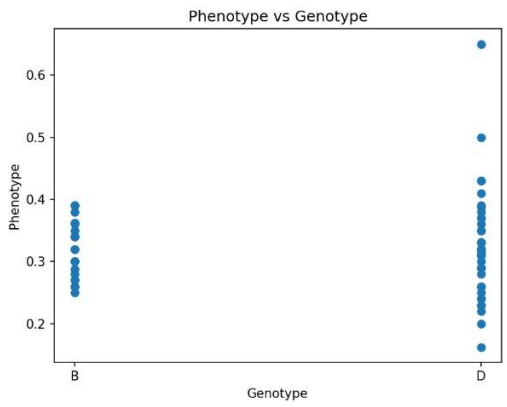
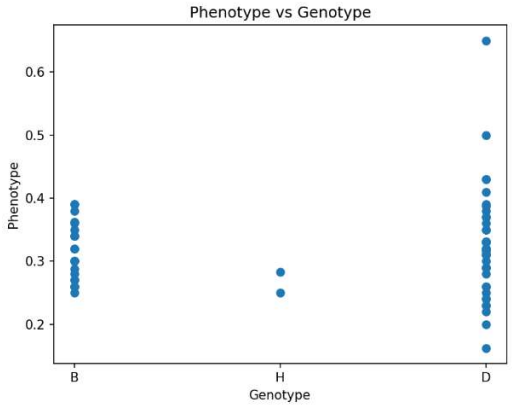
Systems Genetics – Ex2

- 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) Due 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:





a. We consider cases with only B or D genotype.
Given linear regression as we learnt:

$y = \beta_0 + \beta_1 x + \varepsilon$

$such\ that: y\ is\ phenotype\ and\ x\ is\ the\ genotype\ (B\ is\ 0, D\ is\ 1)\ \varepsilon\ is\ the\ error$

$\varepsilon \sim N(0, \sigma^2)$

$Now\ we\ want\ to\ find\ \beta_0\ and\ \beta_1\ , our\ hypothesis\ testing\ will\ be:$

$H_0 => \beta_0 = \beta_1 = 0$

$H_A => \beta_0 \neq \beta_1 \neq 0$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = calc\ in\ python\ (\text{calc\_beta1 func}) = 0.009$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x} = calc\ in\ python\ (\text{calc\_beta0}\ func) = 0.315$$

$$R^2 = 0.0035$$

$To\ calc\ p - value\ we\ can\ use\ F\ test, so\ first\ we\ need\ to\ find$
$the\ F - value\ and\ then\ use\ the\ F\ distibution\ to\ calc\ p - value:$

$$F - value = \frac{\frac{R^2}{(1 - R^2)}}{n - 2} = calc\ in\ python\ (\text{regression\_my\_imp func}) = 0.219$$

$$p - value = P\big(F - value > F_{1,n-2}\big) = 1 - P(F - value < F) = calc\ in\ python = 0.641$$

We get it's insignificant and don't reject H_0 (matches our assumption looking at the data)

$In\ addition\ we\ can\ get\ a\ p - value\ using\ t - test\ regarding\ \beta_1.$
$We\ should\ find\ the\ t\ value\ and\ use\ the\ t\ distribution:$

$$t - val = \frac{\beta_1}{\sqrt{\frac{\frac{sse}{n-2}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} \quad and\ sse = rss = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$calc\ in\ python, we\ get\ \ t - val = 0.468\ \ and\ now\ we\ can\ calc\ the\ p - value$
$p - value = 2 \cdot P(t > t - val) = 2 \cdot (1 - P(t < 0.468) = 0.641$

We get it's insignificant and don't reject H_0
(It matches our assumption by looking at the data)

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

b. This time we consider cases with only B or D or H genotype.
Given linear regression as we learnt:

$y = \beta_0 + \beta_1 x + \varepsilon$
$such\ that: y\ is\ phenotype\ and\ x\ is\ the\ genotype\ (B\ is\ 0, H\ is\ 1\ and\ D\ is\ 2)\ \varepsilon\ is\ the\ error\ (Normal\ distribution)$
$Now\ we\ want\ to\ find\ \beta_0\ \ and\ \beta_1\ , our\ hypothesis\ testing\ will\ be:$
$H_0 => \beta_0 = \beta_1 = 0$
$H_A => \beta_0 \neq \beta_1 \neq 0$
$Similar\ calculations\ as\ before\ lead\ us\ to\ the\ following\ results:$
$\beta_0 = 0.312$
$\beta_1 = 0.005$
$R^2 = 0.0043$
$F - value = 0.28$
$p - value = 0.59$
$\boldsymbol{We\ get\ it's\ insignificant\ and\ don't\ reject\ H_0}$

c. This time we consider only cases with B or D genotype and we are interested to do ANOVA test.
$Using\ what\ we\ leant\ in\ class\ we\ can\ split\ our\ data\ to\ 2\ groups: B\ and\ D$
$The\ model\ suggests:$
$x_{ij} = \mu_i + \varepsilon_{ij}$
$where\ x_{ij}\ is\ the\ phenotype,\ \ \mu_i\ is\ the\ effect\ of\ group\ i, \varepsilon_{ij}\ is\ the\ error$
$The\ null\ hypothesis\ will\ be:$
$x_{ij} = \mu + \varepsilon_{ij}\ \ (the\ effect\ of\ all\ the\ groups\ is\ the\ same)$

$Now\ we\ need\ to\ find\ the\ relevant\ expressions\ for\ ANOVA\ test:$
$SS_{within}, SS_{among}, MS_{within}, MS_{among}$
$Using\ MS_{within}, MS_{among}\ we\ can\ get\ the\ F - value:$
$$F - value = \frac{MS_{among}}{MS_{within}}$$
$and\ then\ find\ the\ p - value = P\left(F - value > F_{df_{among}, df_{within}}\right)$
$The\ implemntation\ in\ python\ is\ in\ anova\_my\_imp\ func.$
$The\ results:$

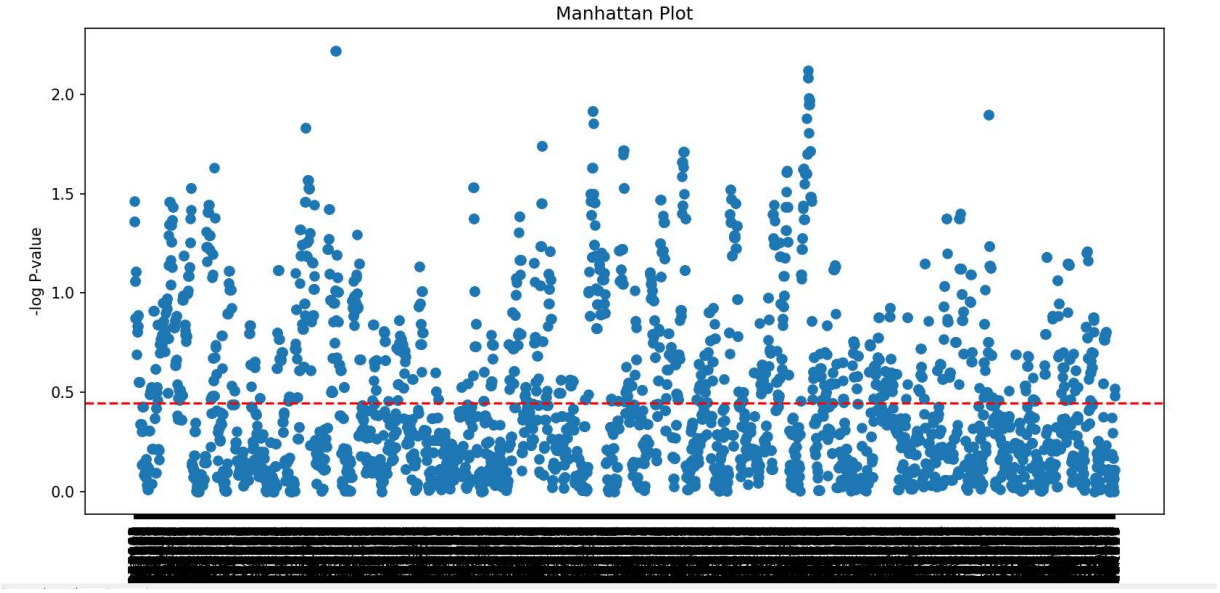
```
F-value: 0.21908676009169 39
p-value: 0.6413779427668764
```

$\boldsymbol{We\ get\ it's\ insignificant\ and\ we\ don't\ reject\ H_0}$
In addition it matches the results of the python stats model ANOVA func 😊

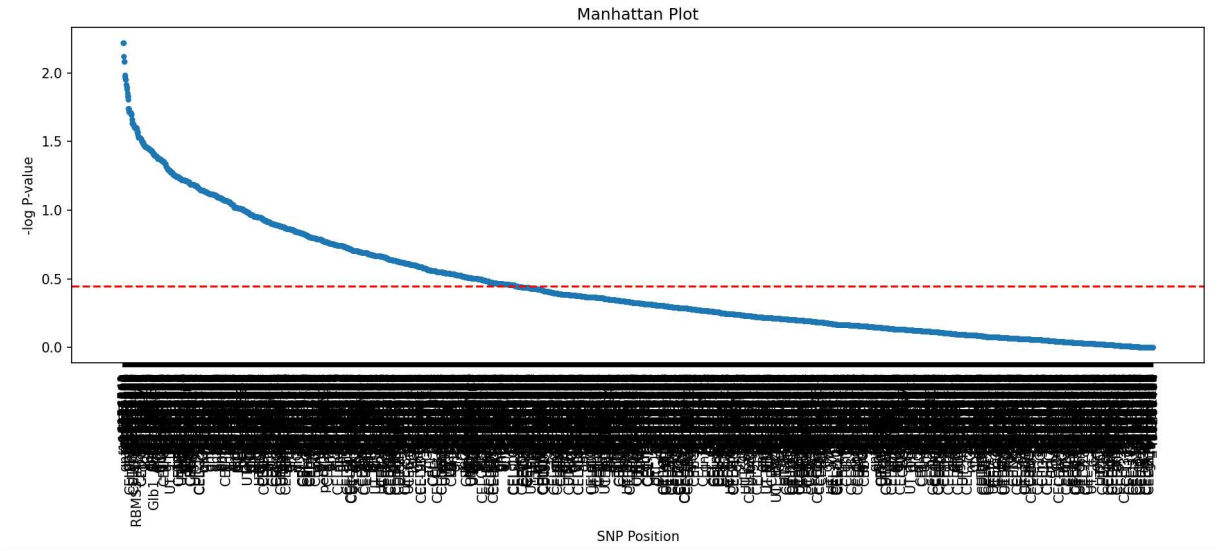
- Comparing the results we got in all tests we can see that (a) and (c) are equivalent and it makes sense by the theory because ANOVA is a generalization to t-test with several groups and here we have only 2 groups so it's similar to regular t-test which we have in the linear regression situation in (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.

2) 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 $1.371 \cdot 10^{-5}$ or in terms of $-\log(p-value) = 4.88$. 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.