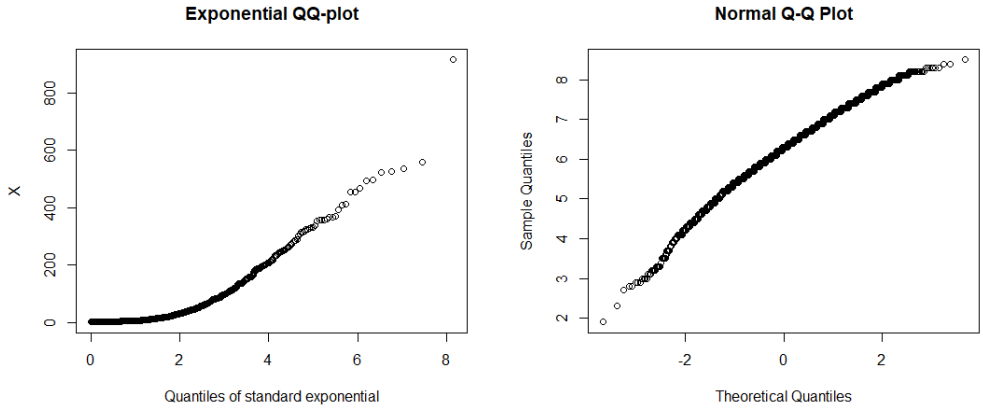
Prerequisite

All code and figures are contained in *https://github.com/feedlord18/CSDS313.git*

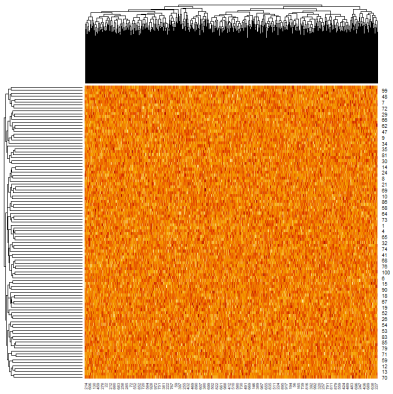
Problem 1

For the airport data-set, in last homework we chose the exponential distribution, therefore we generate a Q-Q plot against a theoretical exponential distribution. For the movie rating data-set, we chose a normal distribution, then we can just plot a Q-Q plot against a theoretical normal distribution.

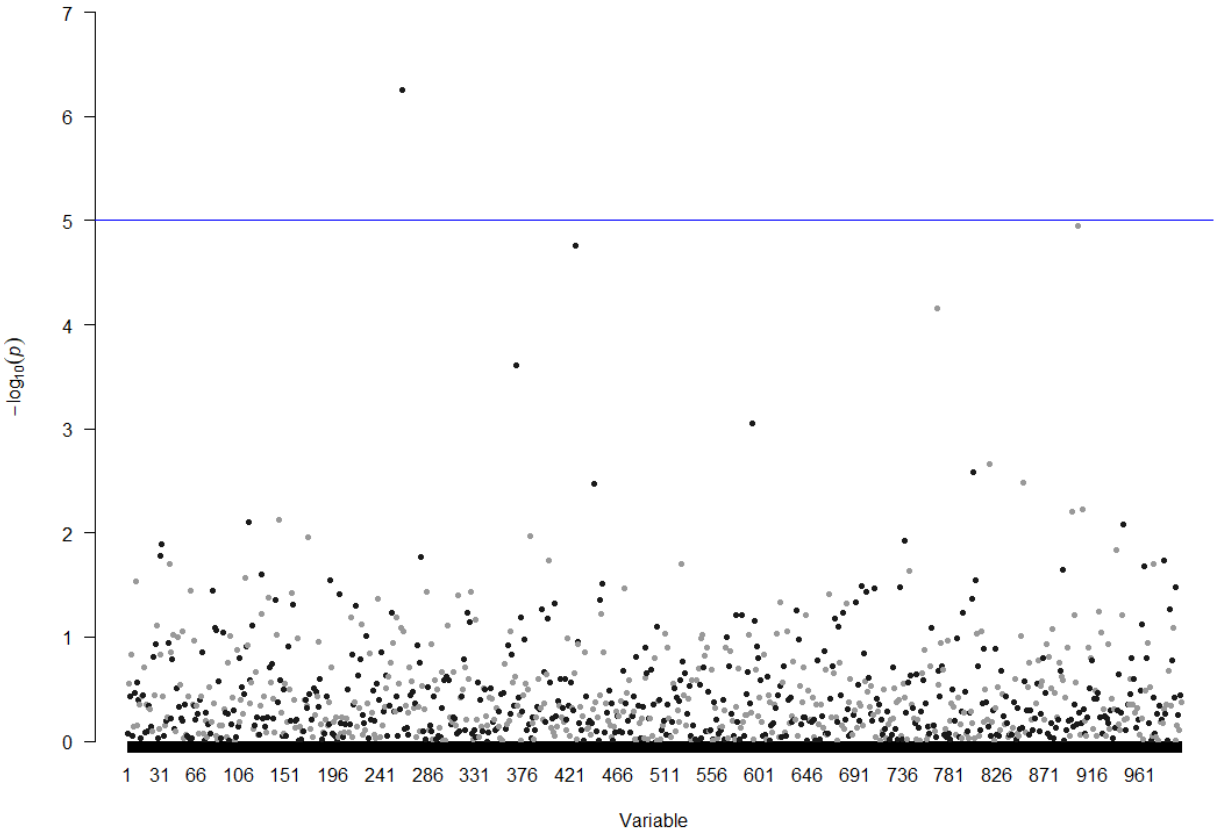


Problem 2

We generated the data and set a constant seed for reproducibility.



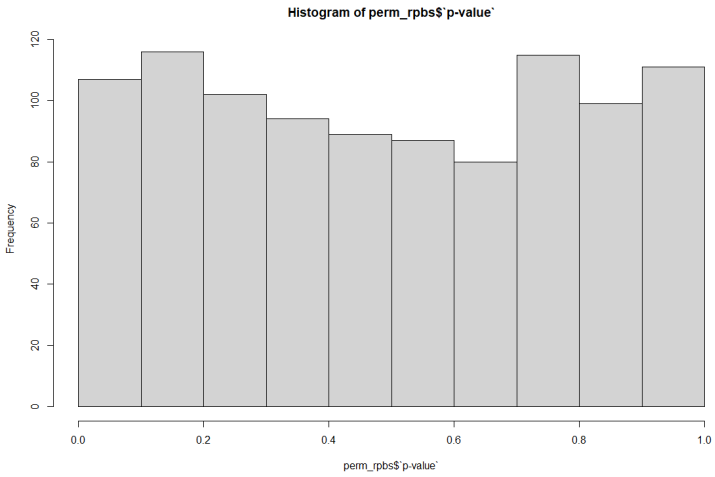
We then can generate a Manhattan map from all the calculated statistics.



Using a significant level of 0.01 reduces to 15 significant variables, which are 115, 144, 261, 369, 425, 443, 593, 768, 803, 818, 850, 896, 902, 906, 945. When applying Bonferroni Correction, we reduce to 1 significant value, which is 261. When using False Discovery Rate, with a q-value of 0.1, we discover 5 significant variables, which are 261, 369, 425, 768, 902. The only value that were all discovered significant which was within the predefined variables is 261. This value is significant in all the test cases because it is one of the five values which has a direct effect on the binary output.

We applied 1000 times of permutation to the binary labels through sampling without replacement and calculated the coefficients again. This time, we notice that there are 18 significant variables given a significant level of 0.01, which are 27, 30, 220, 225, 469, 506, 555, 580, 595, 931. When using a significant level of 0.001, we only obtain 1 significant variable, which is 27. There were no predefined variables which appeared in the significant variables. This is due to sampling over 1000 times, we disassociate the affect of the 5 predefined variables with the outcome. The histogram of the coefficients can be plotted, and we notice a

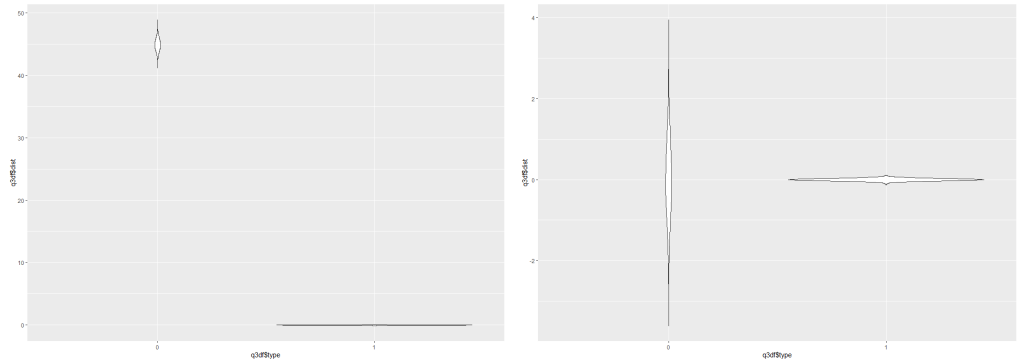
relative even spread among the p-values.



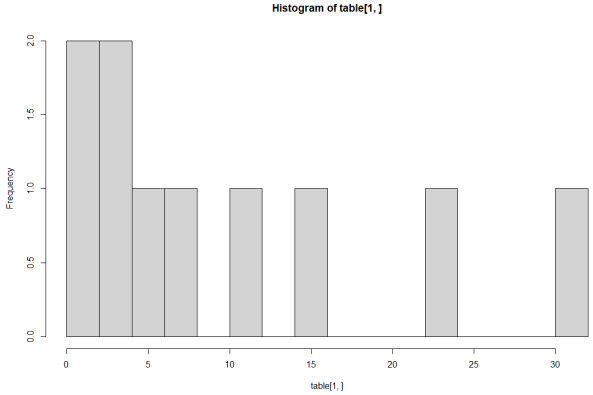
We can see the underlying issues of statistical testing. Since we are trying to determine the contributing variables to the final binary label, we need to understand the structure of the data, in this case, using Point-Biserial Correlation Coefficients we were able to reduce the range of possible contributing factors, by adjusting p-values according, we can reduce or increase the number of statistically significant variables. In the cases where we are not able to determine any significant factors, we can apply numerous other statistical analysis, such as Principle Component Analysis, or some other deterministic algorithms, such as decision trees. We can utilize entropy gains given the variables and see which variables will yield the most information gain compared to the labels.

Problem 3

We can create the violin plot by combining the data points. The plot on the left shows the unprocessed data plot, as euclidean distance have drastic different numeric values than cosine similarity, the plot for cosine similarity is condensed. In order to have a better view, I applied z-normalization since cosine similarity is between -1 and 1, by normalizing the euclidean distance, we can have it centered around 0, which is shown by the figure on the right.



Problem 4



We can notice that when the feature space increases, the distance between samples increase. In other words, we can deduce that once lower dimension data is transposed into higher dimensional space through increase in feature separation, the data distribution becomes more sparse.

	1	2	4	8	16	32	64	128	256	512
Mean	1.184	1.918	2.748	3.852	5.56	7.894	11.240	15.833	22.567	31.787
Variance	0.894	0.976	0.972	0.970	0.959	0.9682	0.985	0.957	0.968	0.945

Table 1: Mean and Variance of Euclidean Distance With Different Feature Space

Problem 5

We can see that since $df = (r-1)(c-1)$, the degree of freedom is 2. The critical value which is standard to chi-squared testing is 0.05. By calculating the statistic, the chi-squared statistic is 13.535, and with a p-value of 0.001156294. Since the p-value is much less than the significant level, we reject H_0 and claim that these two variable are indeed dependent. For

	low	medium	high	sub-sum
For	213	203	182	598
Against	138	110	154	402
sub-sum	351	313	336	1000

Table 2: Expected Counts

DNA mutation data sample, we can perform chi-squared testing and we obtained a p-value of 0.0935. However since there are low frequencies in the contingency table, we should use Fisher's Exact test instead, as it will achieve a better approximation result. Using Fisher's Exact test, we obtain a p-value of 0.0893. Since the p-value of the Fisher's Exact test is higher than the 5% significant level, we accept the null hypothesis which means whether or not the mutation is synonymous or not, is not dependent of whether the mutation is polymorphic or not. We can interpret the p-value of 0.0893 as having a 91.07% chance which the odds ratio will not be equal to 1, and 8.93% chance the odds ratio will equal to 1, where the ratio of synonymous to non-synonymous is equivalent to the odds ratio. Therefore, we can say that the ratio is not the same for polymorphisms and fixed differences.

Question 6

Given the definition of entropy and joint entropy, we can derive the following:

$$H(X) + H(Y \mid X) = -\sum_i p_i \log_2(p_i) - \sum_i \sum_j p_{i,j} \log_2(p_{i,j}/p_i)$$
$$H(X) + H(Y \mid X) = -\sum_i p_i \log_2(p_i) - \sum_i \sum_j p_{i,j} \log_2(p_{i,j}) + \sum_i p_i \log_2(p_i)$$
$$H(X) + H(Y \mid X) = -\sum_i \sum_j p_{i,j} \log_2(p_{i,j})$$
$$H(X) + H(Y \mid X) = H(X, Y)$$

Since the previous is proven, mutual information of X and Y is trivially shown:

$$H(X) + H(Y) - H(X, Y) = H(X) + H(Y) - (H(X) + H(Y \mid X)$$
$$H(X) + H(Y) - H(Y, X) = H(Y) - H(Y \mid X)$$
$$I(X, Y) = H(Y) - H(Y \mid X)$$