The problems on hypothesis testing.

Actually, each of these problems can be solved in a fully digital environment, such as ipynb.
You can submit handwritten solutions as well, but they should be detailed enough to reflect all the important steps.

Model quality [25 points]

I build a new model for music recommendation, and it gave 1235 good recommendations out of 1600. The previous model is known to give 75% of good recommendations. Can I be sure that my model is an improvement? Provide an analytical p-value and a bootstrapped one to test the null hypothesis of no improvement vs the alternative of positive improvement.

Comparing salaries [25 points]

The data are salaries corresponding to two kinds of occupations: (1) creative, media, and marketing and (2) education. Suppose that the datasets are modeled as realizations of normal distributions. Test the null hypothesis that the salary for both occupations is the same at 5% significance level. Don't assume equal variance in two groups.

Counting bombs [25 points]

The table below gives the number of bombs falling into the South of London during WWII. The South of London was divided into n = 576 regions of 0.25 km^2 each. In the table, the number n_k corresponds to the number of domains bombed exactly k times. We want to estimate whether bombs were falling on South of London "at random".

k	0	1	2	3	4	5+
n_k	229	211	93	35	7	1

The total number of bombs is $\sum k n_k = 537$. Can we claim that the number of bombs per region has Poisson distribution? We can answer with the Chi-squared goodness of fit test:

1) Estimate the parameter λ for Poisson distribution from our data, and calculate expected number of regions $\tilde{n}_k$ with number of bombs from 0 to 5 (multiply the corresponding probability by 576).
2) If our assumption of Poisson distribution is true, then n_k has approximately normal distribution with mean $\tilde{n}_k$ and variance $\tilde{n}_k$. If so, the statistic $T = \sum_k \frac{(n_k - \tilde{n}_k)^2}{\tilde{n}_k}$ has approximate χ^2 distribution with 4 degrees of freedom (6 cells, minus 2 restrictions on the counts: total sum and conformity with Poisson distribution). Calculate the sample value of T.
3) Calculate p-value as 1 minus χ^2_4 CDF of T at its sample value. Is it low enough to reject the hypothesis about the Poisson distribution?

Russian cities, continued [25 points]

In the previous home assignment, you worked with the file russian_cities_g_9k.csv to fit the parameters of Pareto distribution. Now, use Kolmogorov-Smirnov method to test whether the data was really generated by the Pareto distributions with the parameters you have found.