Case Study: DS & AI job market exploration on hh.ru Russian vacancy data (2020 vs. 2024)

Alexandra Vabnits, Bulat Akhmatov

https://github.com/sashhhaka/HH-VacancyAnalysis

Datasets (Raw and preprocessed):

case_study - Google Drive
https://drive.google.com/drive/folders/1SuPjf469UKrZ_ixxtDRugw1l1qU48OGO?usp=sharing

Introduction

This study examines Russian job market dynamics, specifically focusing on the hh.ru vacancy pool, which is a mainly Russian job market online platform for finding and posting vacancies. A vacancy pool is a set of job openings available through the platform for a job seeker. Each job listing on the platform is defined by parameters like required experience, suggested salary, job description, etc. The research focuses on data science (DS) and artificial intelligence (AI) jobs, recognizing their status as a comparatively new and promising field, marked by ongoing developments. By delving into the dynamic job market situation, the study aims to offer valuable insights for DS and AI candidates, particularly for students aspiring to enter these evolving fields.

The study uses inflation to accurately adjust salaries for changes in purchasing power over time, ensuring meaningful economic analysis and facilitating more accurate salary comparisons across different years.

The key question addressed is whether the vacancy pool characteristics have changed over the past four years, from 2020 to 2024, considering factors such as salary, required experience, and employer regions.

This investigation provides valuable insights into the shifting landscape of job opportunities, aiding both job seekers and employers in making informed decisions based on current market conditions and requirements. The contextual insight on how the job situation has evolved allows individuals to adapt strategies to the transformed job market, while employers can adjust hiring approaches to align with the current employment landscape.

Data

Data collection

The data for the study consists of 2020 and 2024 datasets. The 2020 dataset contains data about IT vacancies, collected in 2020 by Bersenev *et al* [1].

2024 data was collected through an API request by collecting all the available at the moment of collection (March 2024) non-archived vacancies with a Python parsing script.

Parameters for an API request:

Search string for hh.ru API [2] with keywords:

```
'Data scientist' or 'Data analyst' or 'ML' or 'AI' or 'Machine L
'Artificial Intelligence' or 'Аналитик данных' or 'Data Enginee
'Reinforcement learning' or 'Аналитик-исследователь' or 'Нейросе
'Искусственный интеллект' or 'Машинное обучение'
```

Region code for Russia and all codes connected to Russia (region and city codes): `113`

Data preprocessing

Script for initial data preprocessing.

1. Filter the 2020 dataset from general IT vacancies into specified DS & AI vacancies using the same search string as for the hh.ru API.
2. The 2020 dataset was collected throughout the whole year by separate API accesses, which can be seen by plotting the published number of vacancies by day of the year. In contrast, the 2024 dataset has been collected by a single API request at one point of time, and its published vacancies are distributed on a one-month scale.

   To address this issue, several one-month cuts were taken out of 2020 data, and the one, that had no large gaps between dates was chosen as a representative group to avoid bias due to not consistent parsing of data during the year in the 2020 dataset. 2020 and 2024 one-month distributions of a number of vacancies appeared to have similar visual structures (Fig. 1, Fig. 2).

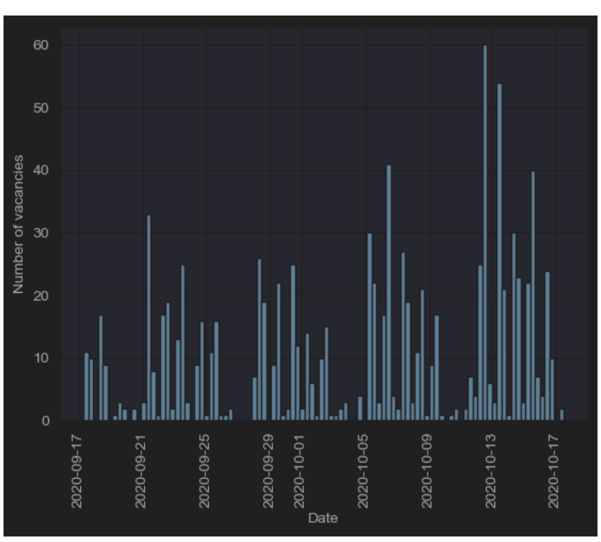

Fig. 1. 2020 data cut by one-month interval and filtered by name

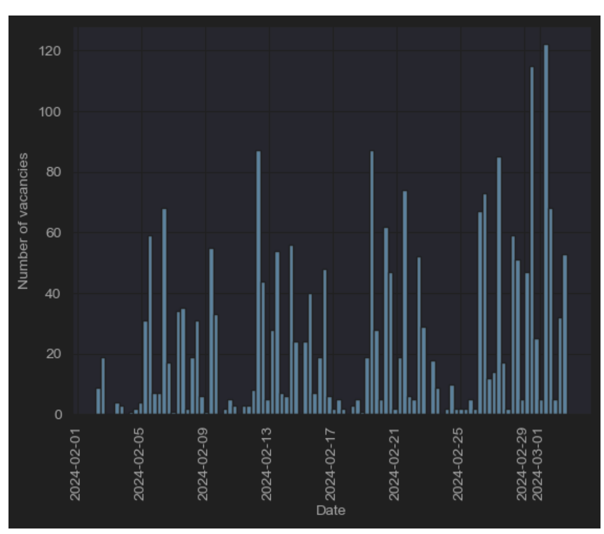

Fig. 2. 2024 raw data

3. Unify dataframe number of columns (2020 - 56 columns, 2024 - 37), leaving only informative ones.
4. Preprocess as needed to have the same data formats. (divide JSON data from 2024 columns into separate) and convert some features to boolean due to excessiveness.
5. Take only Russian vacancies.
6. Account 2020-2024 inflation. To do that, firstly all currencies were translated to RUB (Russian Ruble), then multiplied by the product of monthly inflation (CPI [3]) from 2020.10 to 2024.03, according to official inflation data in Russia [4].

Study protocol

General analysis procedure

- Initial data exploration
- Calculate needed simple statistical parameters
- Data visualization
- Check for normality
- If data is normal, conduct parametrical tests
- Non-parametrical tests
- Resampling if needed

Major steps

The protocol is to conduct the above procedure for comparing 2020 data with 2024 data by features:

1. overall salaries;
2. salaries grouped by required experience;
3. salaries grouped by areas (more specifically, comparing Moscow and St. Petersburg with other cities for both 2020 and 2024)

Note: All salary features were analyzed in terms of lower bound and upper bound separately ("salary from" and "salary to").

Hypothesis testing (list of hypotheses)

Theory & statistical techniques

We have used the Kolmogorov-Smirnov test as a goodness-of-fit test for checking for normality. It turned out that all distributions are not normal, so it was decided to use non-parametric tests.

Non-parametric tests

1. The Kolmogorov-Smirnov Test was chosen to compare distributions because the sample sizes are large enough.
2. Mann-Whitney U Test was chosen to compare distributions with more attention to central tendency compared to kstest.

   Why these tests?
   Wilcoxon signed-rank test doesn't suit, because the data isn't ordinally scaled.
   Kruskal-Wallis H-test doesn't suit, because the test is more commonly used when you have three or more levels. This is why Mann-Whitney was chosen.

Resampling techniques were not used, since all necessary questions were answered on those samples where there was enough sample size.

Statistical tools, other software

- pandas for DataFrame handling, numpy for additional functions calculation
- matplotlib, seaborn for visualization
- from scipy.stats: kstest, mannwhitneyu, zscore for statistical tests and procedures

Results

1. Overall salaries

Column	Number of non-null values	Ratio of non-null values
Salary From 2020	109	0.122
Salary From 2024	345	0.172
Salary To 2020	83	0.093
Salary To 2024	220	0.11

Fig. 2. Available for analysis salary data.

Visualizations

Normalized data in Fig. 3 is for histograms drawn with density=True, which is a more representative form, if we want to see the general frequency distributions. Raw data shows real amount of vacancies per specified salary bound.

Fig. 3. Histogram of number of vacancies per different salaries (lower and upper bounds).

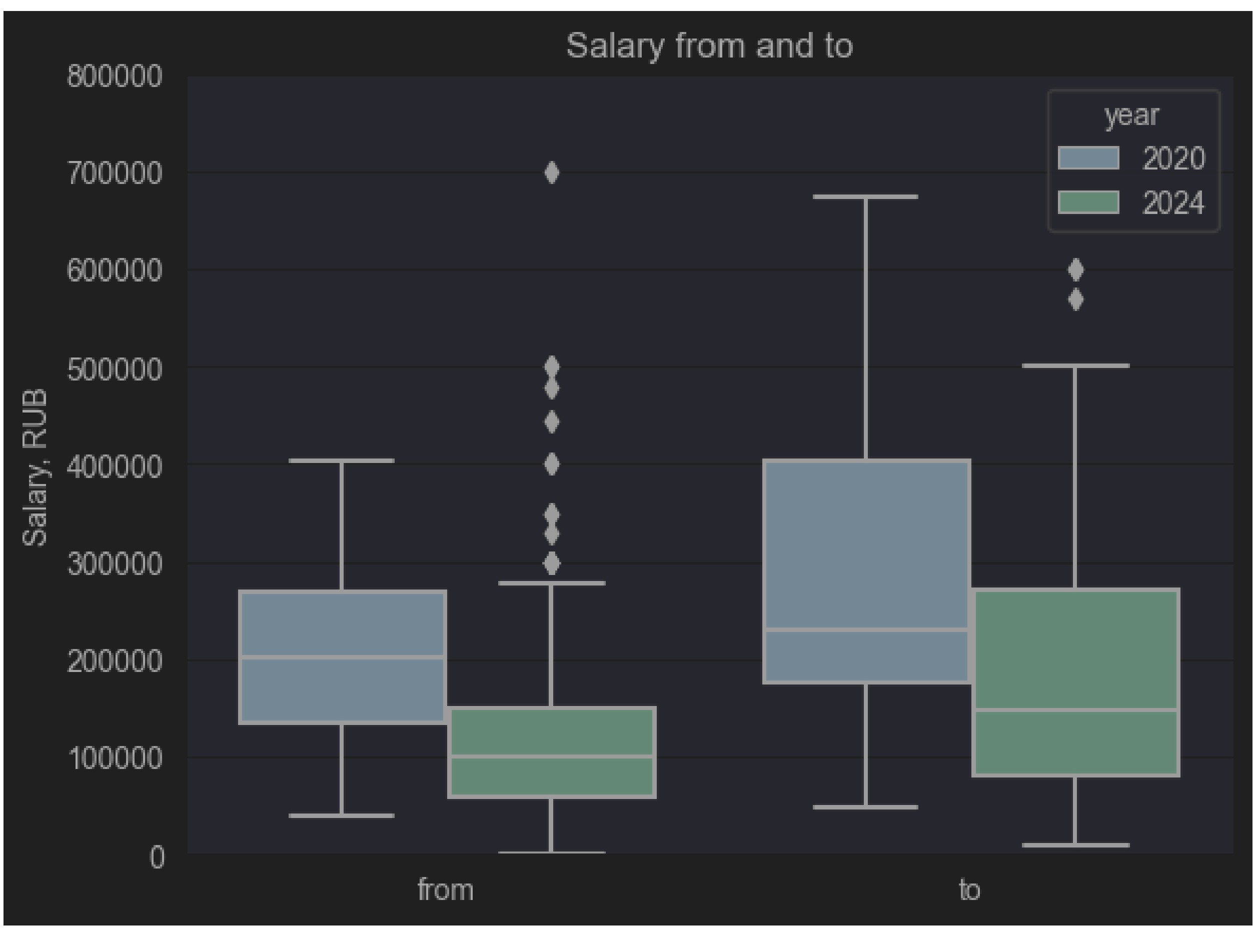

Fig. 4. Boxplot difference between salaries quantiles 2020 vs. 2024

Hypothesis testing

Fig. 5 shows that since the p-values are much less than 0.05, we can reject the null hypothesis of KS test for normality check and conclude that the distributions are not normal. Because of this, we will use non-parametric tests for further analysis.

	Test	KS Statistic	P-Value
0	Salary from 20	0.267606	0.0000
1	Salary from 24	0.185121	0.0000
2	Salary to 20	0.259568	0.0000
3	Salary to 24	0.152430	0.0001

Fig. 5. KS test as a GoF for normality check.

Summary of Statistical Tests

Test	p_value	Stat	Result
Kolmogorov-Smirnov From	9.828036940456279e-18	0.47788857864645656	different
Kolmogorov-Smirnov To	2.6584314675455293e-09	0.40312157721796277	different
Mann-Whitney U From	7.489543387710043e-17	28753.0	different
Mann-Whitney U To	1.0828926210489992e-08	13017.0	different
Mann-Whitney U From One-Sided	3.7447716938550217e-17	8852.0	different
Mann-Whitney U To One-Sided	5.414463105244996e-09	5243.0	different

Fig. 6. Results of all conducted tests for overall salaries.

According to Fig. 6 results, since the p-values are much less than 0.05, we can reject the null hypotheses and conclude that the salaries in 2024 are different from the salaries in 2020 according to both Kolmogorov-Smirnov and Mann-Whitney U tests. Additionally, we can conclude that the salaries in 2024 are less than the salaries in 2020 according to one-tailed Mann-Whitney U tests.

2. Salaries grouped by required experience

Visualizations

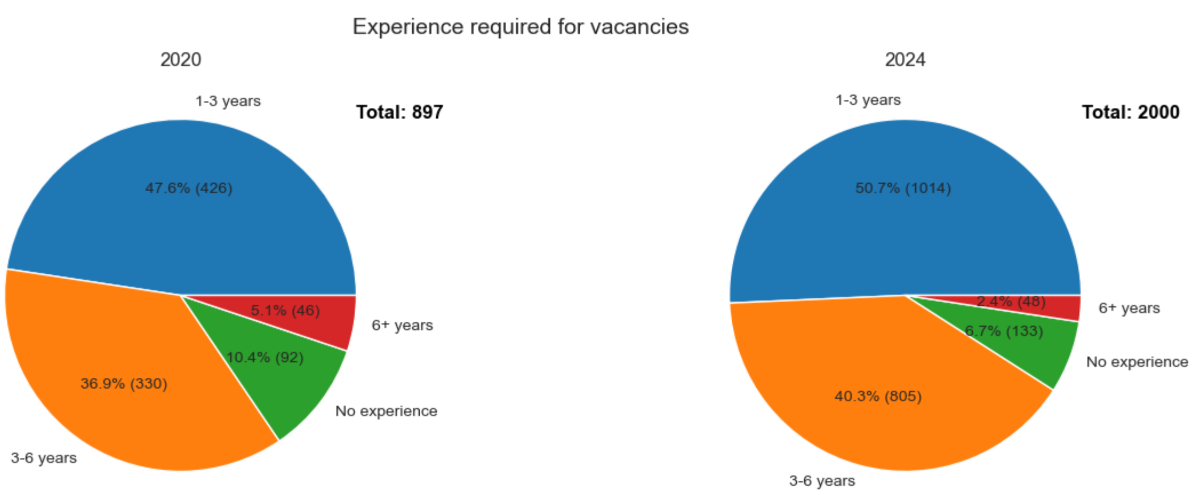

Fig. 7. Fraction of each required experience type in the available vacancy pool

Fig. 7 shows a small decrease in percent of published available vacancies labeled with ‘No experience’ and ‘6+ years’, while ‘1-3 years’ and ‘3-6 years’ fraction of vacancies slightly increased. This change could be further researched with more data published through the whole year. Current research focuses more on change in salaries in each of the groups.

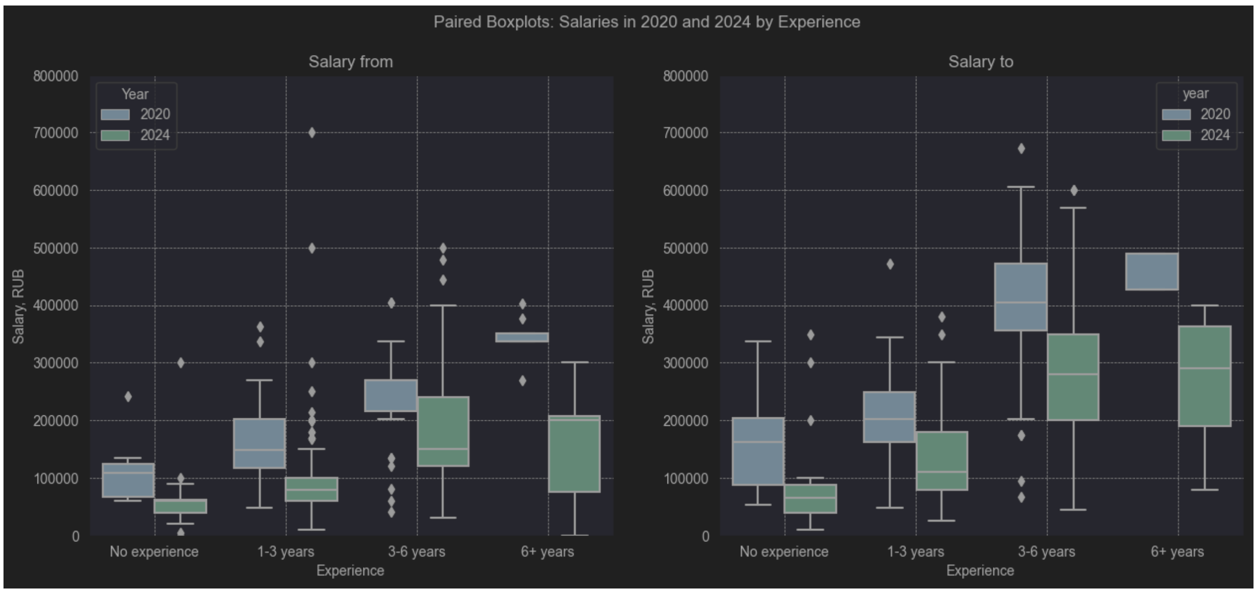

Fig. 8. Boxplots of salaries grouped by required experience

year	experience_id	count	mean	std	min	25%	50%	75%	max
2020	No experience	11.0	108860.028823	51930.428371	60613.851819	67348.724243	107757.958790	124595.139850	2.424554e+05
	1-3 years	50.0	159293.202581	67215.990679	47144.106970	117523.523805	148167.193336	202046.172730	3.636831e+05
	3-6 years	37.0	318796.096324	432257.809093	40409.234546	215515.917579	269394.896974	269394.896974	2.828646e+06
	6+ years	11.0	342791.301179	32488.171028	269394.896974	336743.621217	336743.621217	350213.366066	4.032681e+05
2024	No experience	52.0	58246.057692	39892.489395	50.000000	40000.000000	60000.000000	61250.000000	3.000000e+05
	1-3 years	176.0	96897.945846	71403.683775	10000.000000	60000.000000	80000.000000	100000.000000	7.000000e+05
	3-6 years	105.0	185828.717191	92248.609221	30000.000000	120000.000000	150000.000000	240000.000000	5.000000e+05
	6+ years	12.0	162979.166667	94502.553353	90.000000	75915.000000	200000.000000	207500.000000	3.000000e+05

Fig. 9. Statistical parameters for all experience groups 2020 vs. 2024 year

Fig. 8. and Fig. 9 could show that salaries central tendencies decreased not only overall, but also in each experience group separately. To verify the suppose, hypothesis testing is needed.

Hypothesis testing

The same as for the overall salary changes analysis, we will use the Kolmogorov-Smirnov test and Mann-Whitney U test to compare the salaries in 2020 and 2024 for different experience levels, because the conditions are the same.

H0: the salaries in 2020 are the same as the salaries in 2024;

H1: the salaries in 2020 are different from the salaries in 2024.

Tests results are shown in Fig. 10 and Fig. 11.

Mann-Whitney U Test Results, From:

	Experience Level	p-value	stat	Result
0	6+ years	0.0001	131.0	different
1	1-3 years	0.0	7052.0	different
2	3-6 years	0.0	2842.0	different
3	No experience	0.0	515.0	different

Kolmogorov-Smirnov Test Results, From:

	Experience Level	p-value	stat	Result
0	6+ years	0.0	0.9166666666666666	different
1	1-3 years	0.0	0.5809090909090909	different
2	3-6 years	0.0	0.5521235521235521	different
3	No experience	0.0	0.75	different

Fig. 10. Testing results for lower bound salaries

Mann-Whitney U Test Results, To:

	Experience Level	p-value	stat	Result
0	6+ years	0.0018	60.0	different
1	1-3 years	0.0	3851.0	different
2	3-6 years	0.0014	1140.0	different
3	No experience	0.0116	203.0	different

Kolmogorov-Smirnov Test Results, To:

	Experience Level	p-value	stat	Result
0	6+ years	0.0003	1.0	different
1	1-3 years	0.0	0.5024845955078513	different
2	3-6 years	0.0001	0.5034526051475204	different
3	No experience	0.0405	0.53125	different

Fig. 11. Testing results for lower bound salaries

As we can see, the p-values are much less than 0.05, so we can reject the null hypothesis and conclude that the salaries in 2020 are different from the salaries in 2024 for all experience levels according to both Kolmogorov-Smirnov and Mann-Whitney U tests.

One-tailed Mann-Whitney U test hypothesis to check the direction of changes:

H0: the salaries in 2024 are the same or greater than the salaries in 2020;

H1: the salaries in 2024 are less than the salaries in 2020.

One-tailed Mann-Whitney U Test Results, From:

	Experience Level	p-value	stat	Result
0	6+ years	0.0	1.0	different
1	1-3 years	0.0	1748.0	different
2	3-6 years	0.0	1043.0	different
3	No experience	0.0	57.0	different

One-tailed Mann-Whitney U Test Results, To:

	Experience Level	p-value	stat	Result
0	6+ years	0.0009	0.0	different
1	1-3 years	0.0	1180.0	different
2	3-6 years	0.0007	453.0	different
3	No experience	0.0058	53.0	different

Fig. 12. One-tailed Mann-Whitney U test results for separate experience groups.

According to results in Fig. 12, we can conclude that the salaries in 2024 are less than the salaries in 2020 for all experience level groups according to one-tailed Mann-Whitney U tests.

3. Salaries grouped by areas

More specifically, comparing Moscow and St. Petersburg with other cities for both 2020 and 2024.

In this section, "salary" means "salary from" (lower bound for salary), since we decided to not overload notebook with both “from” and “to” salaries.

Visualization

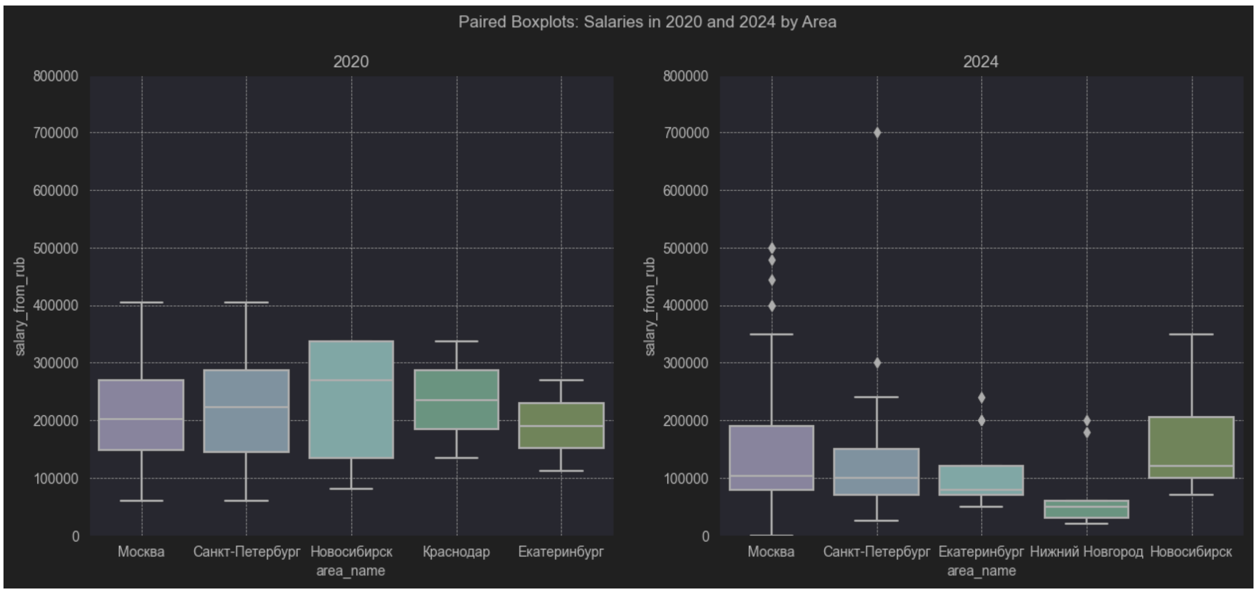

Fig. 12, salaries of top 5 areas (cities) by the number of vacancies.

```
Москва              52
Санкт-Петербург     24
Новосибирск          5
Самара               3
Пермь                3
Name: area_name, dtype: int64
Москва              157
Санкт-Петербург      46
Екатеринбург         13
Новосибирск          11
Владивосток          10
Name: area_name, dtype: int64
```

Fig. 13, number of vacancies for top 5 areas by the number of vacancies.

While the fig. 12 shows that there is some visually noticeable difference in salaries (especially for 2024 year) for different areas, fig. 13 shows that we can say that data is acceptably relevant only for Moscow (Москва) and St.Petersburg (Санкт-Петербург), because for other number of vacancies is under 20, which may be considered insufficient.

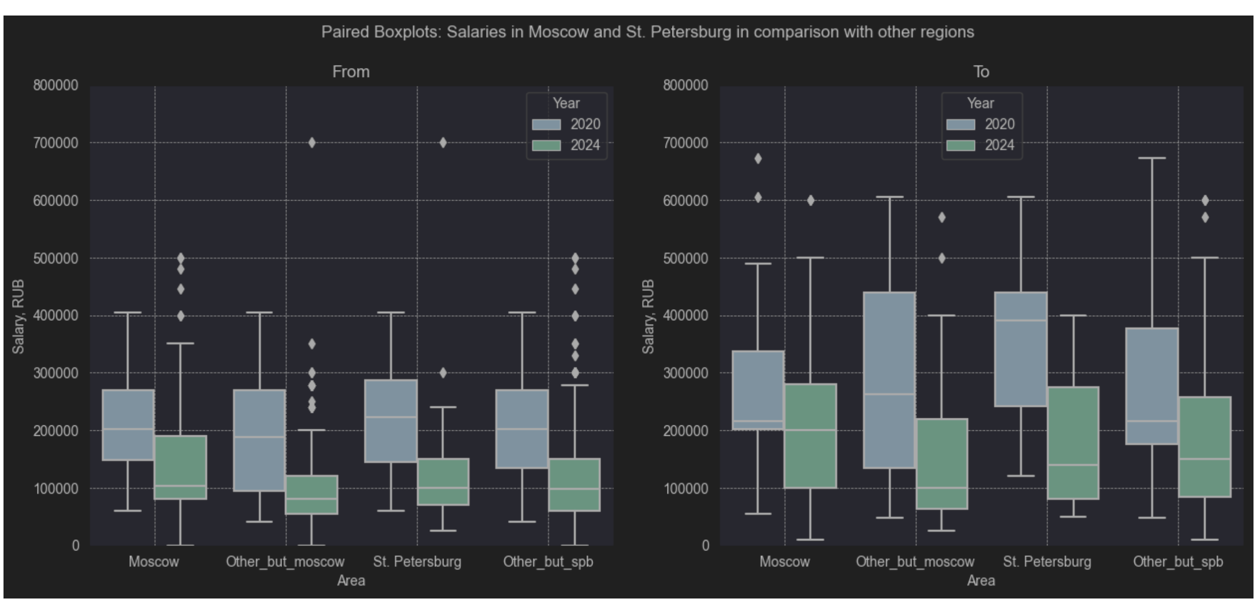

Fig. 14, salaries in Moscow and St. Petersburg in comparison with other regions.

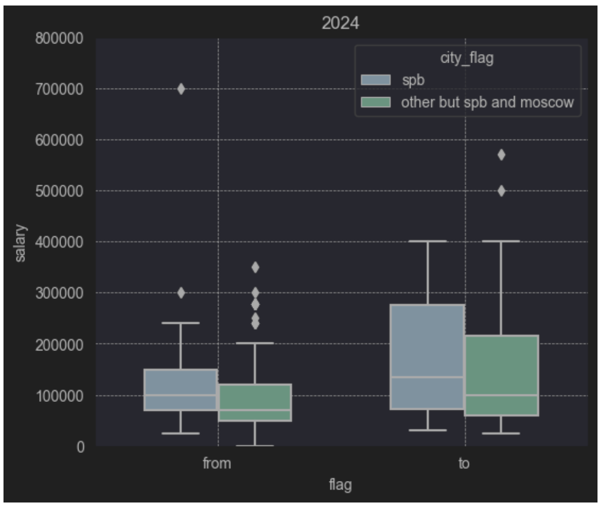

Fig. 15, salaries in St. Petersburg in comparison with other regions excluding Moscow and St.Petersburg

Hypothesis testing

We conducted Mann-Whitney U test for Москва and Санкт-Петербург, and areas except Москва, areas except Санкт-Петербург for 2020 and 2024 years with:

1. H0: the salaries in Москва are the same as the salaries in other areas;

H1: the salaries in Москва are different from the salaries in other areas.

2. H0: the salaries in Санкт-Петербург are the same as the salaries in other areas;

H1: the salaries in Санкт-Петербург are different from the salaries in other area

Mann-Whitney U Test Results

	Year	Region	p-value
0	2020	Москва	0.1892769880609525
1	2020	Санкт-Петербург	0.24154138293724892
2	2024	Москва	8.702692136345948e-07
3	2024	Санкт-Петербург	0.352962002443592

Fig. 16, result of Mann-Whitney U Test for Москва and Санкт-Петербург, and areas except Москва, areas except Санкт-Петербург

In fig. 16, we can see that the p-value is less than 0.05 only for Moscow in 2024 year, so we can reject the null hypothesis and conclude that the salaries in Moscow are different from the salaries in other regions for 2024 year.

There is one problem appears from this fact: since the number of vacancies from Moscow is much higher than from other regions, the salaries from Moscow have a significant impact on the overall distribution of salaries, so it should be better to compare the salaries in St. Petersburg with the salaries in other regions but Moscow and St. Petersburg, to see better the difference in salaries for Saint Petersburg as "the second capital" of Russia.

```
2024: p-value for St. Petersburg vs. other regions: 0.0027
```

Fig. 16, result Mann-Whitney U Test for Санкт-Петербург, and areas except Москва and Санкт-Петербург

Fig. 17 shows that the p-value is less than 0.05, so we can reject the null hypothesis and conclude that the salaries in St. Petersburg are different from the salaries in other areas, excluding Moscow and St. Petersburg

Conclusion

Answers

All conclusions, of course, about hh.ru for DS specialists.

1. Salaries in 2024 have decreased compared to 2020 when adjusted for inflation, indicating a downward trend in compensation for Data Science positions.
2. Across all levels of required experience (ranging from no experience to over six years), salaries in 2024 are lower than those in 2020, suggesting a general decline in compensation regardless of seniority.
3. In 2020, there was no significant difference in salaries between Moscow and other regions, as well as between St. Petersburg and other regions. However, by 2024, salaries in Moscow have diverged from those in other regions.

Additional findings:

1. The distribution of salaries is not normal.
2. The salaries in St. Petersburg are different from the salaries in other areas, excluding Moscow and St. Petersburg.

In summary, the vacancy pool characteristics for Data Science and Machine Learning positions on hh.ru have significantly changed over the past four years (period from 10.20 to 0.2.24), with notable changes in salary levels and regional disparities, along with a non-normal distribution of salaries highlighting potential anomalies within the dataset. These findings provide valuable insights for stakeholders in the field of Data Science recruitment and compensation.

Addressing biases

Firstly, hh.ru does not show all the available vacancies on the job market, we have the access only to a fraction of public online available data and do not know what vacancies may be suggested inside companies or by private invitation. Results of the study applies only to public vacancies.

Secondly, we do not have the detailed procedure of collection of the 2020 dataset. Although we have undertaken measures to ensure preprocessed datasets for 2020 and 2024 are similar in structure, there still may be nuances, such as api search strings that yields non comparable sets of vacancies.

Additionally, usage of consumer price index as inflation measurement method may be not the most suitable method, it was used as measure recommended by POCCTAT. Further economical research is needed to address this bias. Also to be noted, in the 3 hypothesis testing, St. Petersburg had 24 data points, which may be considered as insufficient by some resources.

References

[1] Aleksandr Bersenev, Andrey Sozykin, Denis Shadrin, Anton Koshelev, Evgeniy Kuklin, Alexander Aksenov, March 4, 2021, "IT vacancies from hh.ru, 2006-2020", IEEE Dataport, doi: https://dx.doi.org/10.21227/6naz-wb22.

[2] hh.ru API: https://dev.hh.ru/

[3] Consumer price pndex description: https://en.wikipedia.org/wiki/Consumer_price_index

[4] POCCTAT site with official data about inflation in Russia: https://rosstat.gov.ru/statistics/price

Contributions of co-authors

Alexandra Vabnits: data retrieval, 1-2 research steps, report.

Bulat Akhmatov: data preprocessing, 2-3 research steps, report.