Miranda_Zhao_AnalyticsChallenge

2022-10-03

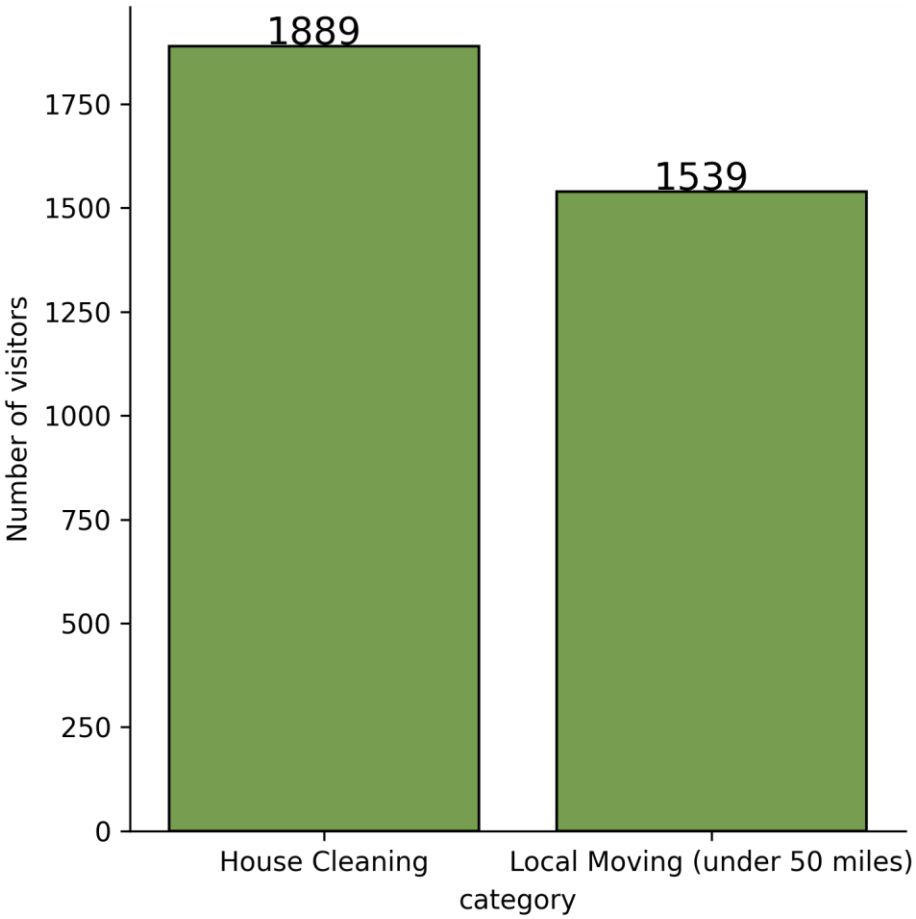
Introduction

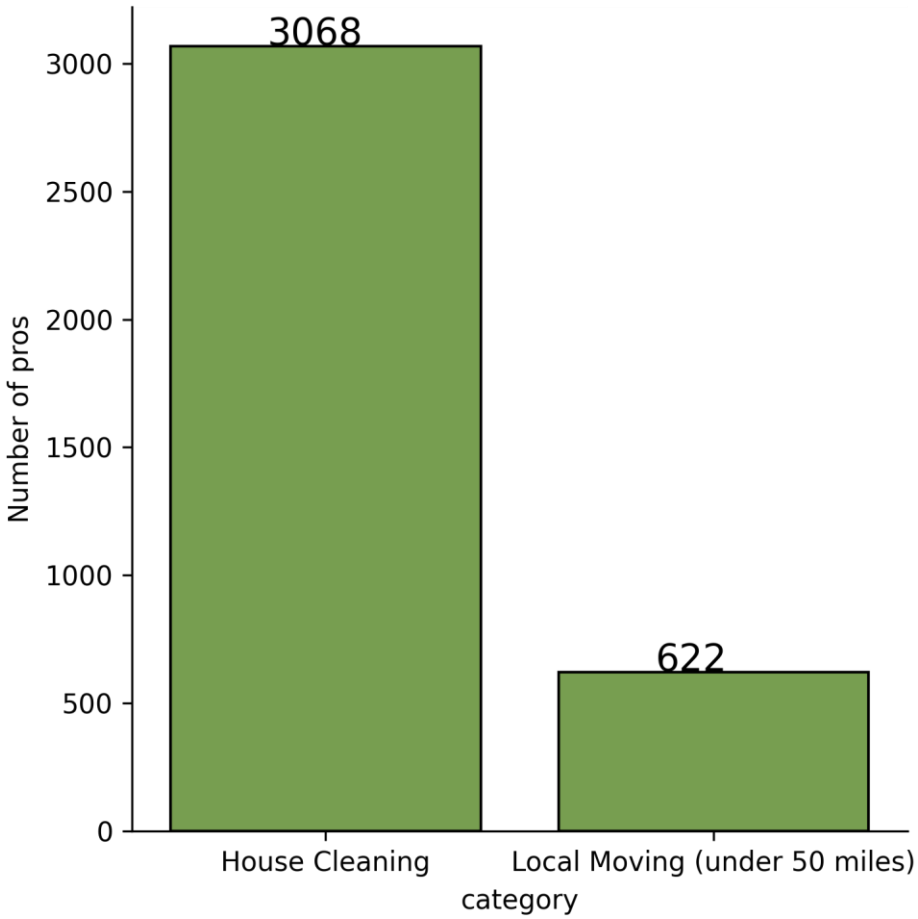
Thumbstack is a platform for people to hire pros. Customers can provide basic information about their projects and search for pros and see their pricing. From the list of pros returned by a search, customers can then view pro's profiles and service pages, contact the pros that interest them, and ultimately hire a pro. Thumbstack generates its revenue by charging pros a lead generation fee for each customer that contacts them.

As a future analyst at Thumbstack, I analyzed a collection of search 26,102 results, representing 3,662 pros returned by searches from 3,428 customers between 2018-10-01 and 2018-11-30, along with customers' actions, namely, viewing pro's profile (or not), contacting the pro (or not), and hiring the pro (or not). One pro can show up in multiple searches by different customers. By analyzing these search and hiring data, I want to find out what types of pros customers are interested in. This report contains my discoveries and data-driven recommendations for improving and growing our marketplace.

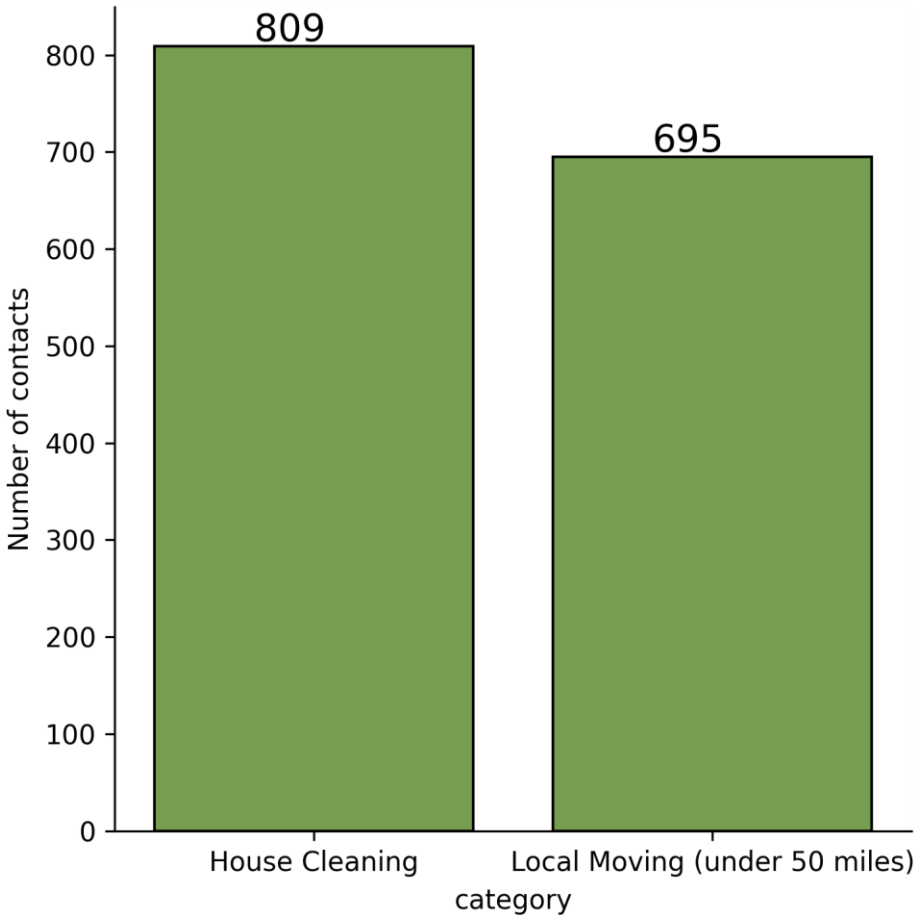
Understanding the Big Picture

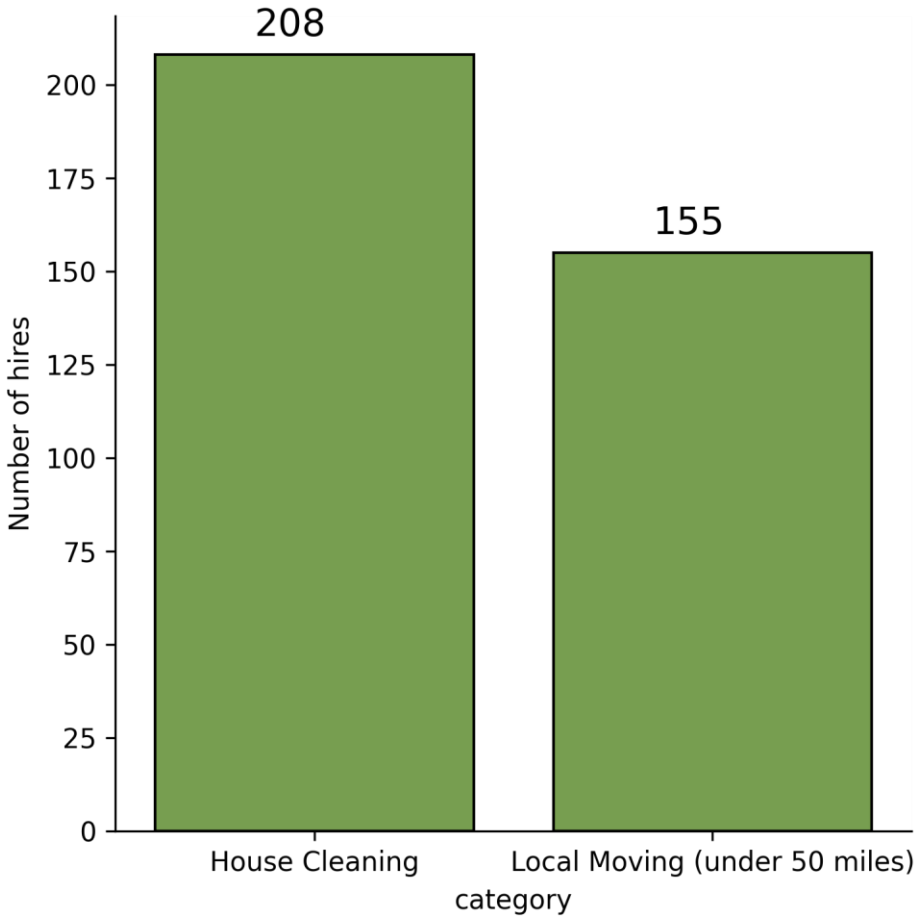
The dataset contains search results for Thumbstack's two major categories, namely, house cleaning and local moving (under 50 miles). There are slightly more people seeking help for house cleaning than local moving, however, the number of pros who offer house cleaning services is 50x of those who offer local moving services.


1889
1539
1750
1500
1250
1000
750
500
250
0
Number of visitors
House Cleaning
Local Moving (under 50 miles)
category



The dataset contains 1,504 contacts and 363 hires in total over two months. This translates to an annual average of 4.9 contacts per pro and 2.9 hires per contact. Assuming these rates could hold up well over time, then adding more pros to our platform would result an increase in contacts and hires, which would in turn increase the revenue. On the other hand, it implies that a typical customer, on average, makes 5.3 contacts per year and 1.3 hires per year. If we could find ways to increase these rates, we would be able to increase contacts and hires and hence revenue without needing to significantly increase the number of customers. The following figure breaks down the number of contacts and hires for house cleaning and local moving respectively. Of the pros contacted by customers looking for house cleaning, 25.7% got hired, and of those who were contacted for local moving, 22.3% got hired.


809
695
800
700
600
500
400
300
200
100
0
Number of contacts
House Cleaning
Local Moving (under 50 miles)
category

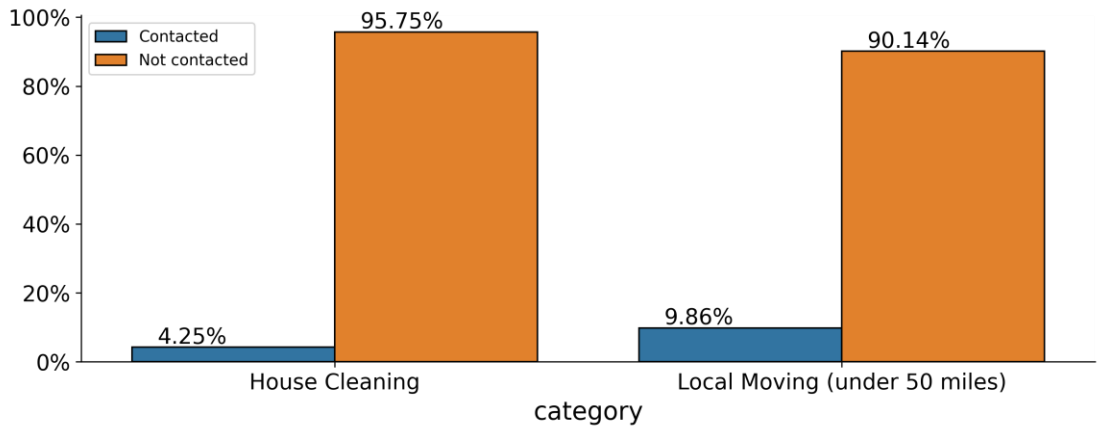


What types of pro get contacted?

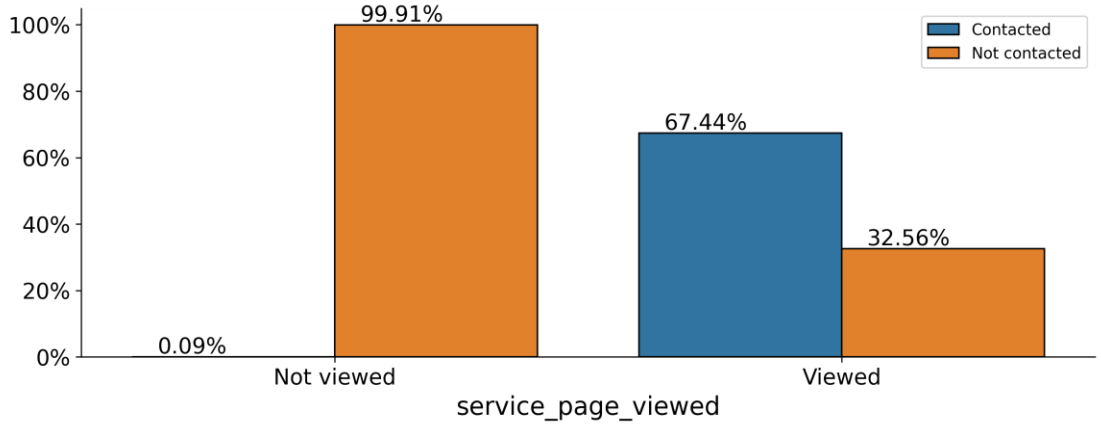
Thumbtack generates revenue by charging pros for each customer that contacts them. So by understanding the traits of those pros who got contacted, we could design strategies to promote these traits and ultimately increase sales. Let's start with some charts calculated and plotted at the 26,102 search results level instead of unique customer level or pro level.

Exploratory Analysis

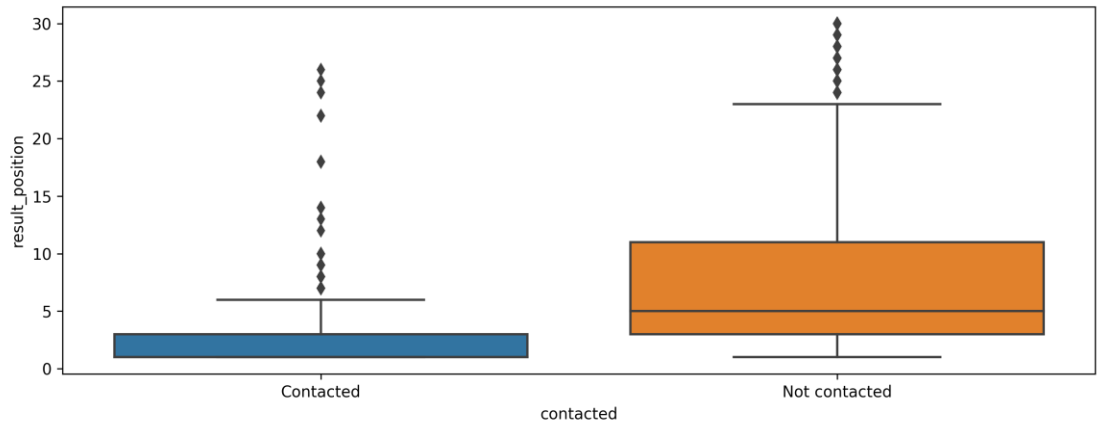
The contact rate is 4% ~ 10% for house cleaning and local moving respectively.



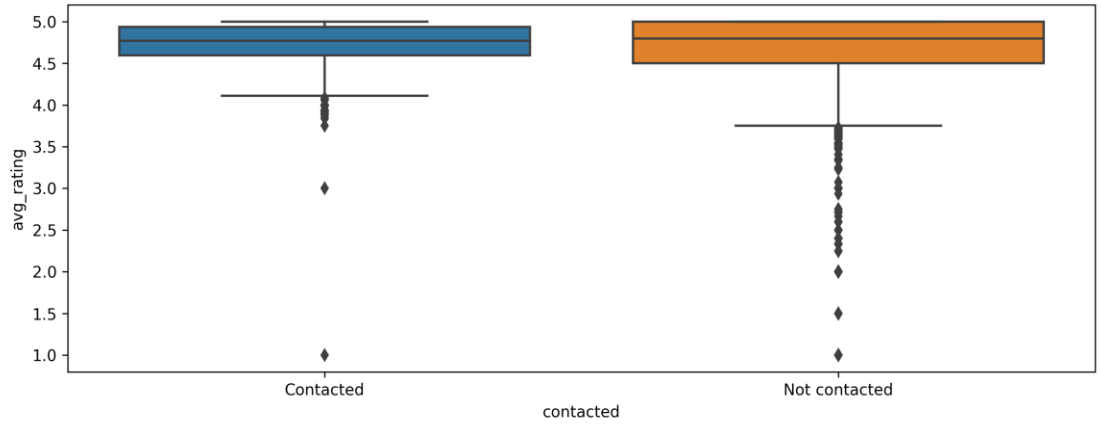
How often do customers contact the pro after viewing his or her service page? The answer is 67.4% (chart below). Conversely, when customers don't visit pro's profile or service page, customers almost never hires that pro.



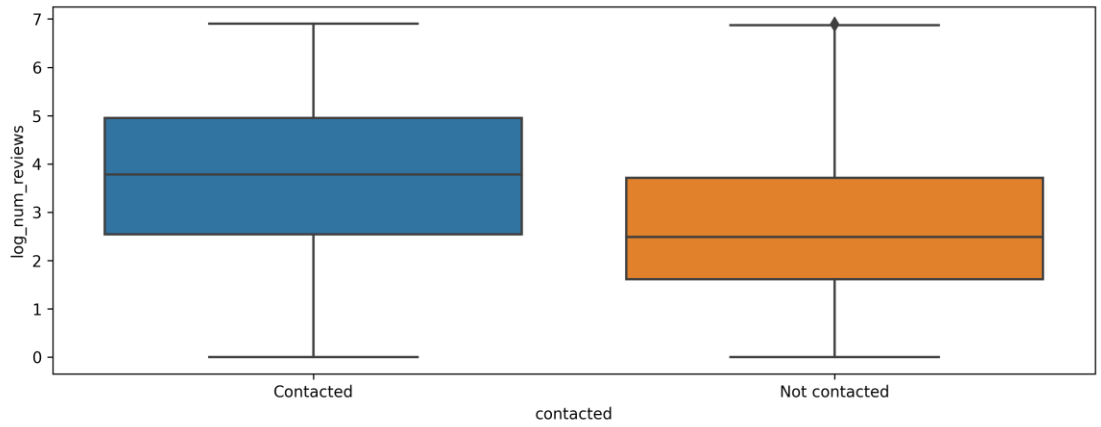
Does pro's rank in search results affect if the pro gets contacted? The answer is a definite yes as seen from the following chart. Most of the pros that got contacted (50% of the data, blue band) had ranks below 4 or 3, where rank 1 is the best. On the other hand, most of the pros that weren't contacted by customers (50% of the data, orange band) had ranks ranging from 4 to 10. It's clear that customers prefer to contact pros ranked top 3 in search results.



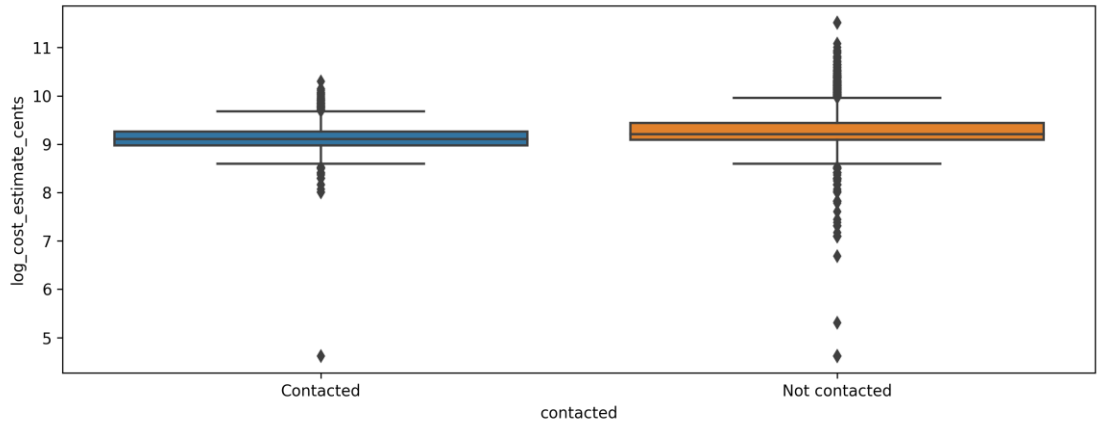
What about pro's average rating? Does it influence if the pro gets contacted? The answer is also yes as we can see from the following figure that pros that were contacted by customers have way fewer outliers below rating 4 than those that weren't contacted. The former group also has 50% of its data concentrated between 4.5 and 5 (blue band), whereas the latter group has its 25th percentile slightly below 4.5 (orange band).



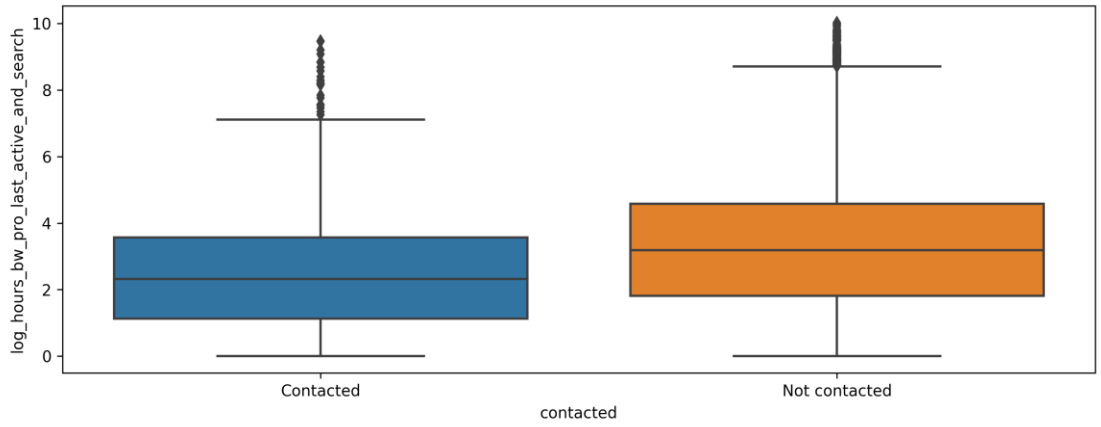
The number of reviews also affects if the pro gets contacted. The following chart shows the number of reviews in log scale for pros that were contacted by customers and those weren't. It's clear that most of those that got contacted had more reviews. The blue band is taller than the orange band.



Pro's pricing doesn't seem to affect if the pro gets contacted. In the following chart, those pros that were skipped by customers have a much wider range in their pricing with more outliers in both directions. The orange band is only tiny bit taller than the blue band.



The number of hours between pro's last time active and customer search seems to also affect if the pro would be contacted by the customer. In the chart below, we see the orange band sits taller than the blue band, implying less active pros got ignored and more active ones got contacted.



Finally, the following table shows the mean values of pro's traits, broken down by contact status, category, and service page view status. Pro's rank in search results, number of reviews, and last active time are highly differentiable, making them potential drivers for customers' decision on whether to contact a pro. Pro's average rating and pricing are not differentiable across different groups and hence are non-drivers.

contacted	category	service_page_viewed	result_position	avg_rating	log_num_reviews	log_cost_estimate_cents	log_hours_bw_pro_last_active_and_search
Contacted	House Cleaning	Not viewed	3.83	4.80	3.31	9.26	2.92
		Viewed	2.61	4.77	3.20	9.19	2.85
	Local Moving (under 50 miles)	Not viewed	5.00	4.82	3.10	8.99	2.84
		Viewed	1.76	4.68	4.32	9.07	2.10
Not contacted	House Cleaning	Not viewed	8.95	4.68	2.42	9.32	3.62
		Viewed	3.42	4.71	3.11	9.21	3.02
	Local Moving (under 50 miles)	Not viewed	4.10	4.62	3.32	9.20	2.79
		Viewed	2.20	4.67	4.20	9.09	2.39

Logistic Regression Analysis

Now that we gained some understanding from exploratory analysis, let's fit a logistic regression to the data to rigorously study how each of pro's traits affects customer's decision on whether to contact a pro or not.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.4544	1.599	-4.661	0.000	-10.589	-4.319
result_position	-0.1697	0.026	-6.523	0.000	-0.221	-0.119
avg_rating	0.4918	0.190	2.593	0.010	0.120	0.864
log_num_reviews	0.0207	0.039	0.525	0.600	-0.056	0.098
log_cost_estimate_cents	-0.1347	0.143	-0.939	0.347	-0.416	0.146
log_hours_bw_pro_last_active_and_search	-0.0500	0.029	-1.698	0.089	-0.108	0.008
local_moving	0.4149	0.106	3.919	0.000	0.207	0.622
Viewed	7.3740	0.257	28.733	0.000	6.871	7.877

Looking at the above table of coefficient estimates, we see that pro's rank in search results, average rating, service page viewed status, and category are highly statistically significant with p-value < 0.001, and last time active is almost significant with a p-value of 0.089. On the other hand, the (log) number of reviews and (log) cost estimate are not statistically significant. The following table lists the odds ratios derived from the coefficient estiamtes.

	OR	Lower CI	Upper CI
Intercept	0.00	0.00	0.01
result_position	0.84	0.80	0.89
avg_rating	1.64	1.13	2.37
log_num_reviews	1.02	0.95	1.10
log_cost_estimate_cents	0.87	0.66	1.16
log_hours_bw_pro_last_active_and_search	0.95	0.90	1.01
local_moving	1.51	1.23	1.86
Viewed	1,594.00	963.90	2,635.98

- Pro's rank in search results has an odds ratio of 0.84, meaning that for 1 unit decrease in rank, (remember smaller rank is better,) we can expect a 16% increase in the odds of being contacted.
- Average rating has an odds ratio of 1.64, meaning that for 1 unit increase in average rating, we can expect a 64% increase in the odds of being contacted.
- Because the (log) number of reviews, (log) cost estimate, and pro's last active time are not statistically significant, the meaning of their odds ratios cannot be generalized beyond the sample data. So we'll ignore them here.
- Local moving has an odds ratio of 1.51, meaning that we can expect the odds of being contacted for local moving pros is 51% more than house cleaning pros.

Recommendations

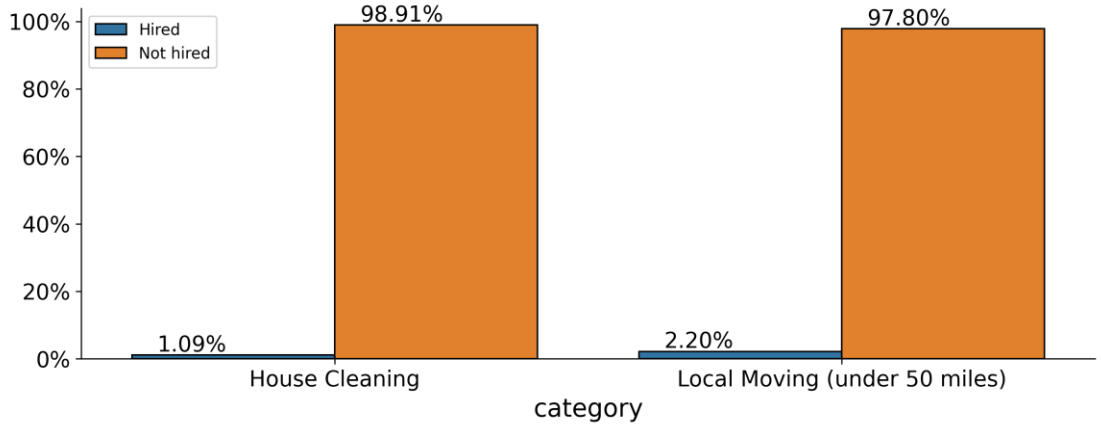
Based on the above exploratory and logistic regression analyses, I recommend we only show top 10 pros in search results because customers seem not to look beyond 10. We also want to show pros with higher average ratings first or give them special status such as top rated plus. Finally, local moving services are highly demanded but short supplied, so we want to get more local moving pros on to the platform. We may have to spend some marketing money to acquire them, but this expense would be well justified given that simply being a local moving pro would increase the odds of being contacted and hence revenue generation by 51%.

What types of pro get hired?

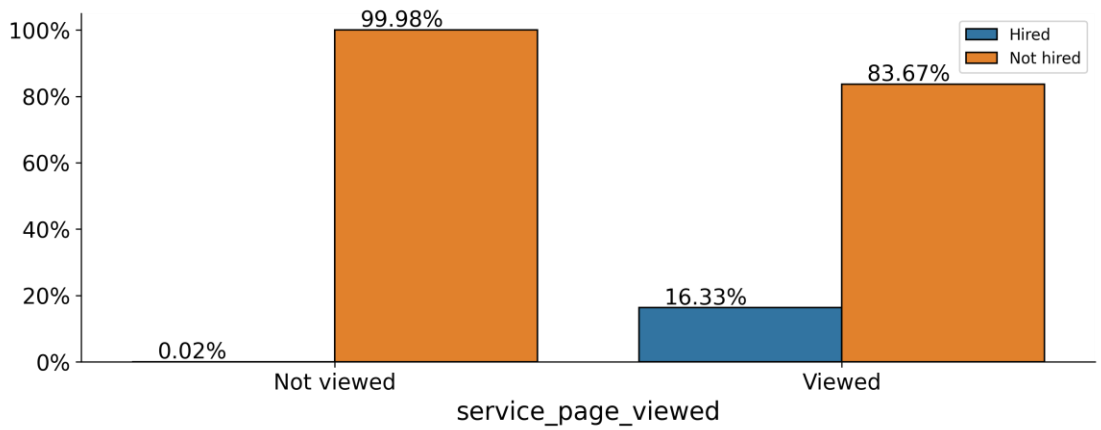
Having understood the things that get customers to contact a pro, let's further examine the things that get a pro hired. Once again, let's start with some charts.

Exploratory Analysis

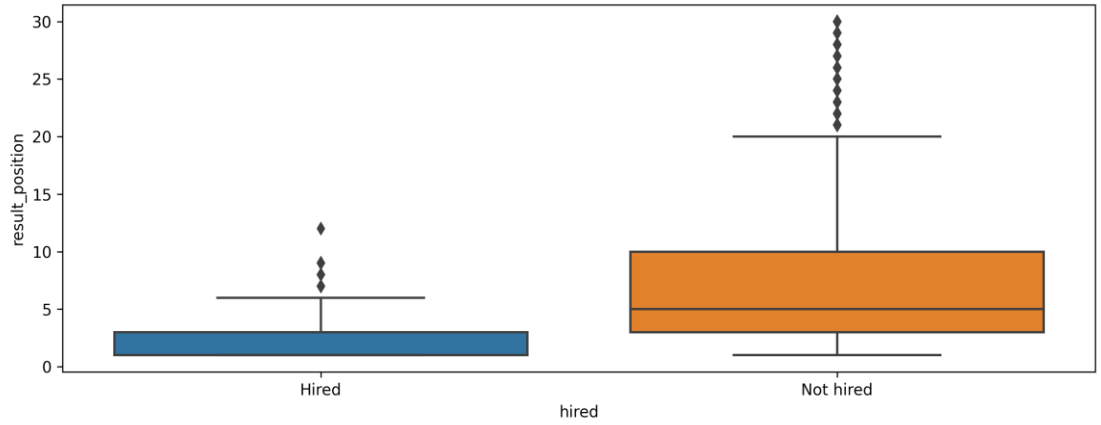
The hiring rate is 1% ~ 2.2% for house cleaning and local moving.



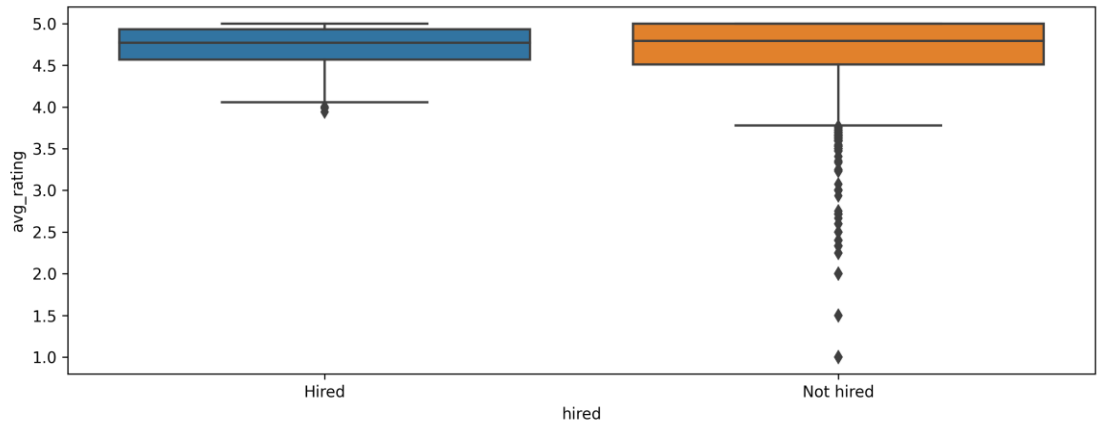
Does hiring tend to follow after viewing a pro's service page? The answer is yes as we can see from the following chart that service page viewing corresponds to 16.3% hiring whereas not viewing has almost 0% hiring.



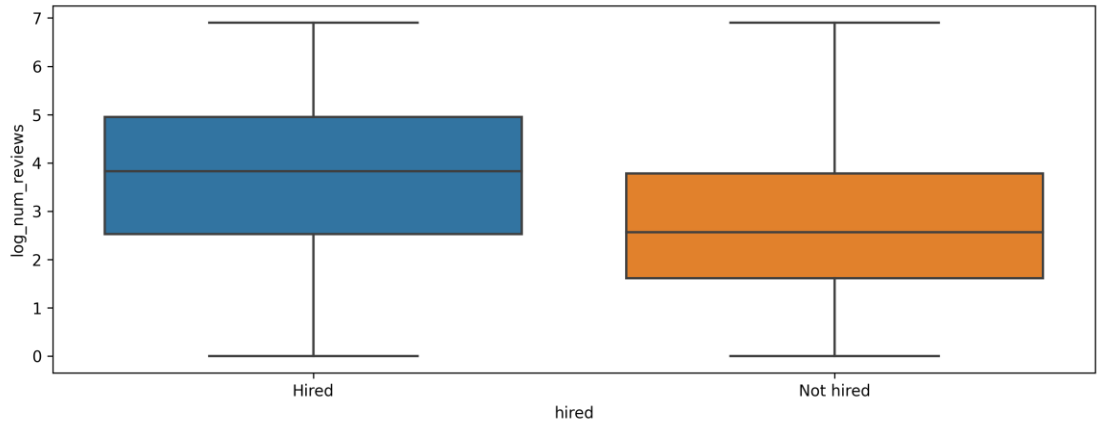
Does pro's rank in search results affect if the pro gets hired? The answer is a definite yes. The chart below shows that most of the pros that got hired (50% of the data, blue band) had ranks below 4 or 3, with rank 1 being the best. On the other hand, most of the pros that weren't hired (50% of the data, orange band) had ranks ranging from 4 to 10. It's clear that customers prefer to hire pros ranked top 3 in search results.



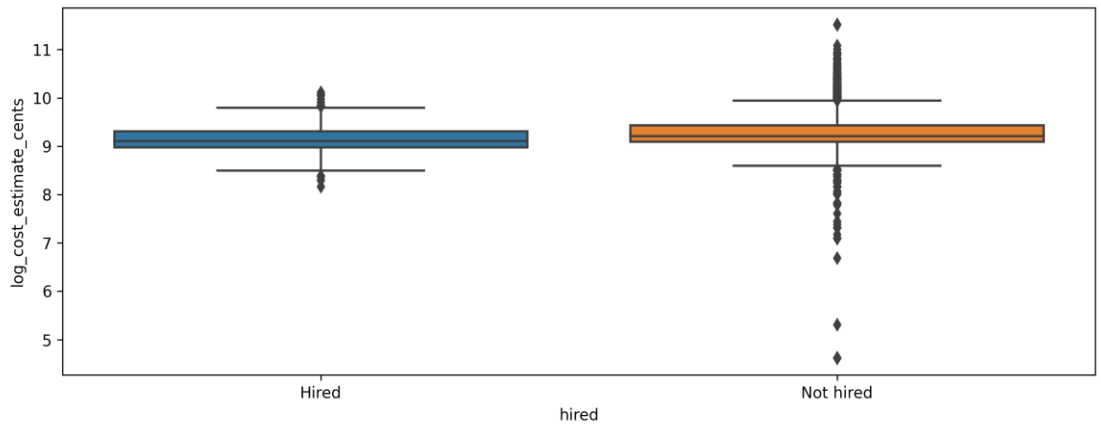
What about pro's average rating? Does it influence if the pro gets hired? The answer is also yes. Pros that got hired have way fewer outliers below rating 4 than those that didn't. Pros that were hired has 50% of its data concentrated between 4.5 and 5 (blue band), whereas those that weren't hired has its 25th percentile slightly below 4.5 (orange band).



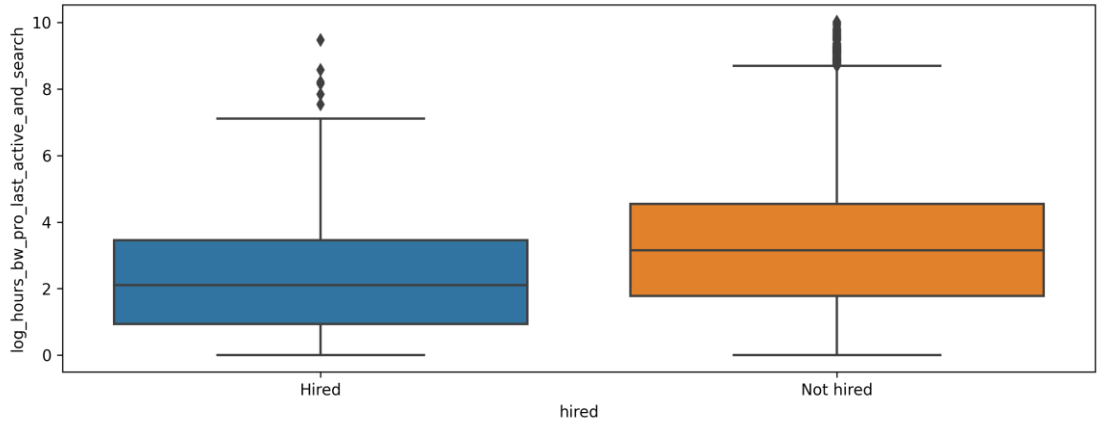
The number of reviews also affects if the pro gets hired. The following chart shows the number of reviews in log scale for pros that were hired by customers and those weren't. It's clear that most of those that got hired had more reviews. The blue band is taller than the orange band.



Pro's pricing doesn't seem to affect if the pro gets hired. In the following chart, pros that weren't hired have a much wider range in their pricing with more outliers in both directions. The orange band is only tiny bit taller than the blue band.



The number of hours between pro's last time active and customer search seems to also affect if the pro would be hired by the customer. In the chart below, we see the orange band sits taller than the blue band, implying less active pros got passed and more active ones got hired.



Finally, the following table shows the mean values of pro's traits, broken down by hire status, category, and service page view status. Pro's rank in search results, number of reviews, and last active time are highly differentiable, making them potential drivers for customers' hiring decision. Pro's average rating and pricing are not differentiable across different groups and hence are non-drivers.

hired	category	service_page_viewed	result_position	avg_rating	log_num_reviews	log_cost_estimate_cents	log_hours_bw_pro_last_active_and_search
Hired	House Cleaning	Not viewed	1.75	4.68	4.62	9.33	2.57
		Viewed	2.40	4.78	3.21	9.21	2.64
	Local Moving (under 50 miles)	Viewed	1.99	4.66	4.39	9.05	2.13
Not hired	House Cleaning	Not viewed	8.95	4.68	2.42	9.32	3.62
		Viewed	3.01	4.74	3.15	9.19	2.97
	Local Moving (under 50 miles)	Not viewed	4.10	4.62	3.32	9.20	2.79
		Viewed	1.85	4.68	4.27	9.08	2.18

Logistic Regression Analysis

Now that we gained some understanding from exploratory analysis, let's fit a logistic regression to the data to rigorously study how each of pro's traits affects customer's hiring decision.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-9.0556	2.199	-4.119	0.000	-13.365	-4.746
result_position	-0.0963	0.039	-2.497	0.013	-0.172	-0.021
avg_rating	0.1954	0.263	0.742	0.458	-0.321	0.712
log_num_reviews	0.0017	0.050	0.035	0.972	-0.096	0.099
log_cost_estimate_cents	0.0101	0.188	0.053	0.957	-0.359	0.379
log_hours_bw_pro_last_active_and_search	-0.0756	0.039	-1.936	0.053	-0.152	0.001
local_moving	-0.0525	0.132	-0.396	0.692	-0.312	0.207
Viewed	6.7978	0.587	11.572	0.000	5.646	7.949

Looking at the above table of coefficient estimates, we see that pro's rank in search results, last active time, and service page viewed status are statistically significant with p-value < 0.05. And pro's average rating, the (log) number of reviews, the (log) cost estimate, and category are not statistically significant. The following table lists the odds ratios derived from the coefficient estiamtes.

	OR	Lower CI	Upper CI
Intercept	0.00	0.00	0.01
result_position	0.91	0.84	0.98
avg_rating	1.22	0.73	2.04
log_num_reviews	1.00	0.91	1.10
log_cost_estimate_cents	1.01	0.70	1.46
log_hours_bw_pro_last_active_and_search	0.93	0.86	1.00
local_moving	0.95	0.73	1.23
Viewed	895.87	283.28	2,833.18

- Pro's rank in search results has an odds ratio of 0.91, meaning that for 1 unit decrease in rank, (remember smaller rank is better,) we can expect a 9% increase in the odds of being hired.
- The (log) hours between pro's last active time and customer's search has an odds ratio of 0.93, meaning that for 2.72 hours increase, we can expect a 7% decrease in the odds of being hired. So customers prefer to hire pros that are recently active.
- Service page view status has a huge odds ratio of 895, meaning that there's almost zero chance for a pro to be hired without first being viewed by customers. This makes sense as this is a direct consequence of how the platform works.
- Because average rating, the (log) number of reviews, (log) cost estimate, and category are not statistically significant, the meaning of their odds ratios cannot be generalized beyond the sample data. So we'll ignore their interpretations.

Recommendations

Based on the above exploratory and logistic regression analyses, I recommend that we show only the top 10 pros in search results because customers seem not to look beyond 10.

This will also lessen the load of the platform and improve page response time. We also want to show pros with higher average ratings first and give them special status such as top rated plus. Finally, we want to incentivize pros to be active all the time and signal their last active times to make it easy for customers to see.

Code:

https://colab.research.google.com/drive/1TeL9WxXE17GaMmSEHwwf1ZFSvd5OvgYE#scrollTo=9-CUvQ-1hS2S