Data Analyst : Activity

Activity 1: There are two attached files which has dummy data for some activity by contractors on the App. File titled Approved Missions has the complete list of Approved Missions in a (random) 30 day period, along with earnings per mission. Note - a contractor only gets paid for an 'approved' mission. A mission goes into 're-attempt' if he fails a quality check. A 'rejected' mission is when he does not complete the mission within the time limit OR makes too many errors, and loses a life on SquadRun. The second file (titled SR Player Details) has demographic data on the contractors including how much they have earned till date (including the data captured for the month), one way in which they are evaluated on the platform (Quality Score = Mean of (2*Approved_Count - 2*Rejected_Count - 1*Re-attempt_Count)) etc.

Can you create segments of contractors based on activity and derive any interesting insights about these segments? The segments can be basis quality, earnings, volume of missions etc.

Please find the required data set here:
•Approved Missions
•Player Details - SR

Activity 2: *This* data set has results from Tests we ran on the platform. There are also results from various missions which have been running on the platform. Based on available data, try to arrive at a framework and/or obtain insights into the performance of a 'Skilled Contractor' and an 'Unskilled Contractor'.

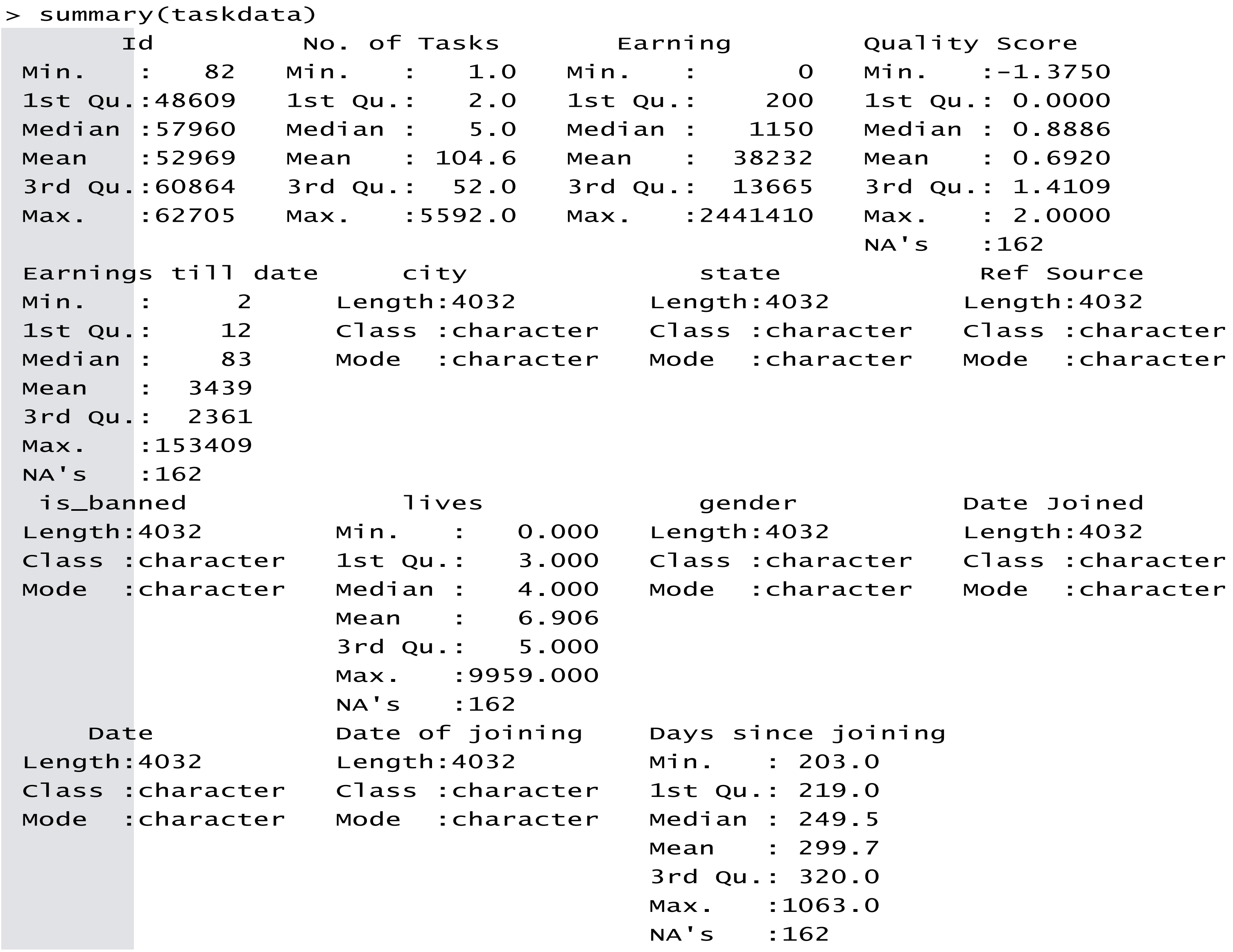
Elaborate on any one particular metric/characteristic that you would be of particular benefit to us in increasing contractor productivity. Explain why you would choose this particular performance/characteristic/metric.

This is an open ended activity set and you are free to take any approach that makes sense to you. Please take care to explore and explain the methodology you adopt in detail.

Note: Visualizations are important. Questions are welcome!

Task 1

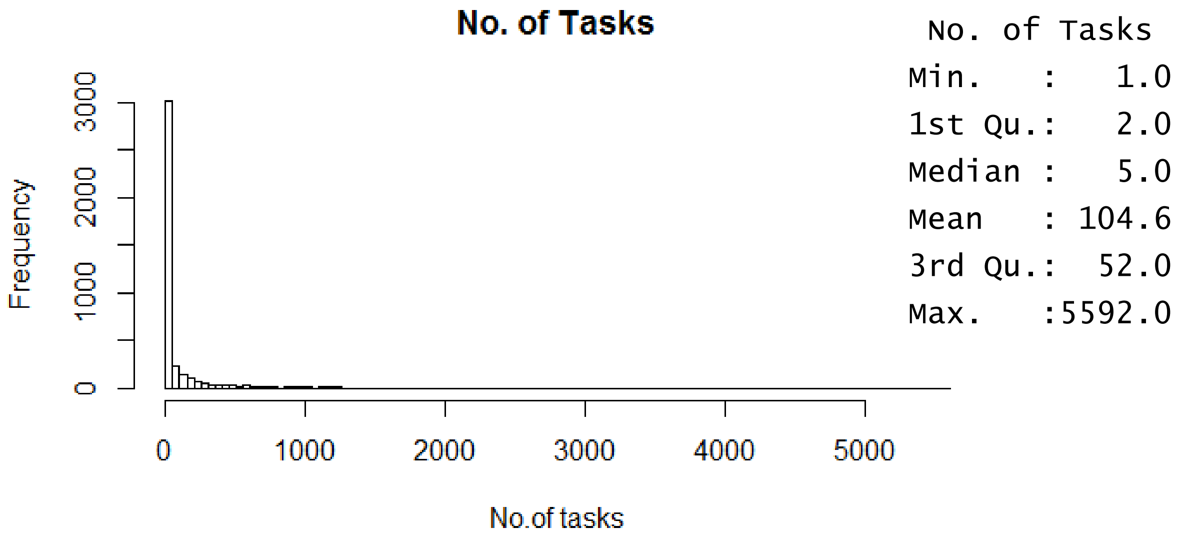
Summary of Data Shared for participants active in that month



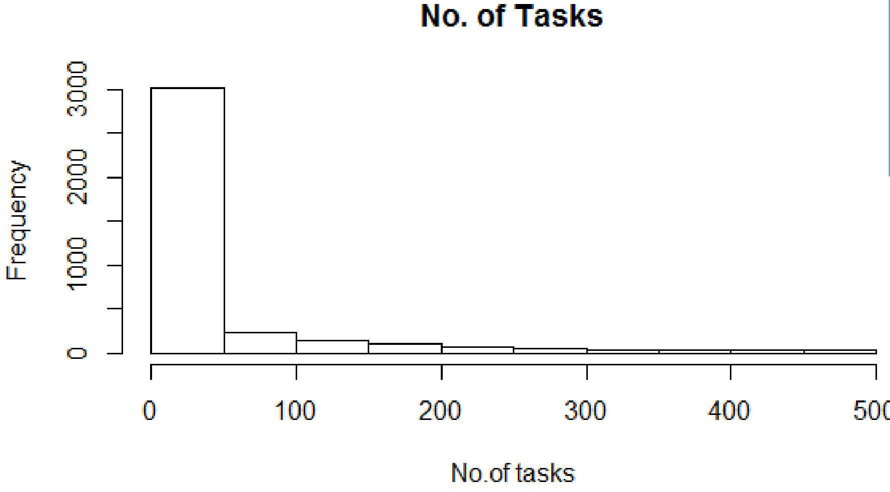
```
> summary(taskdata)
       Id         No. of Tasks        Earning          Quality Score
 Min.   :   82   Min.   :   1.0   Min.   :      0   Min.   :-1.3750
 1st Qu.:48609   1st Qu.:   2.0   1st Qu.:    200   1st Qu.: 0.0000
 Median :57960   Median :   5.0   Median :   1150   Median : 0.8886
 Mean   :52969   Mean   : 104.6   Mean   :  38232   Mean   : 0.6920
 3rd Qu.:60864   3rd Qu.:  52.0   3rd Qu.:  13665   3rd Qu.: 1.4109
 Max.   :62705   Max.   :5592.0   Max.   :2441410   Max.   : 2.0000
                                                    NA's   :162
 Earnings till date      city              state            Ref Source
 Min.   :     2     Length:4032        Length:4032        Length:4032
 1st Qu.:    12     Class :character   Class :character   Class :character
 Median :    83     Mode  :character   Mode  :character   Mode  :character
 Mean   :  3439
 3rd Qu.:  2361
 Max.   :153409
 NA's   :162
  is_banned             lives              gender          Date Joined
 Length:4032        Min.   :   0.000   Length:4032        Length:4032
 Class :character   1st Qu.:   3.000   Class :character   Class :character
 Mode  :character   Median :   4.000   Mode  :character   Mode  :character
                    Mean   :   6.906
                    3rd Qu.:   5.000
                    Max.   :9959.000
                    NA's   :162
     Date           Date of joining    Days since joining
 Length:4032        Length:4032        Min.   : 203.0
 Class :character   Class :character   1st Qu.: 219.0
 Mode  :character   Mode  :character   Median : 249.5
                                       Mean   : 299.7
                                       3rd Qu.: 320.0
                                       Max.   :1063.0
                                       NA's   :162
```

Label	Significance
Id	Id of the player/ vendor
No. of Tasks	No. of tasks attempted in the 30 days period
Earnings	Earnings in the 30 days period
Quality Score	Overall quality score of the player/vendor
Earnings till date	Sum of total earning till date
City	Demographic details
State	Demographic details
Ref Source	As indicated by title
Is banned	False for active
Lives	No. of lives left
Days since joining	Calculated taking 22-5-2017 as reference

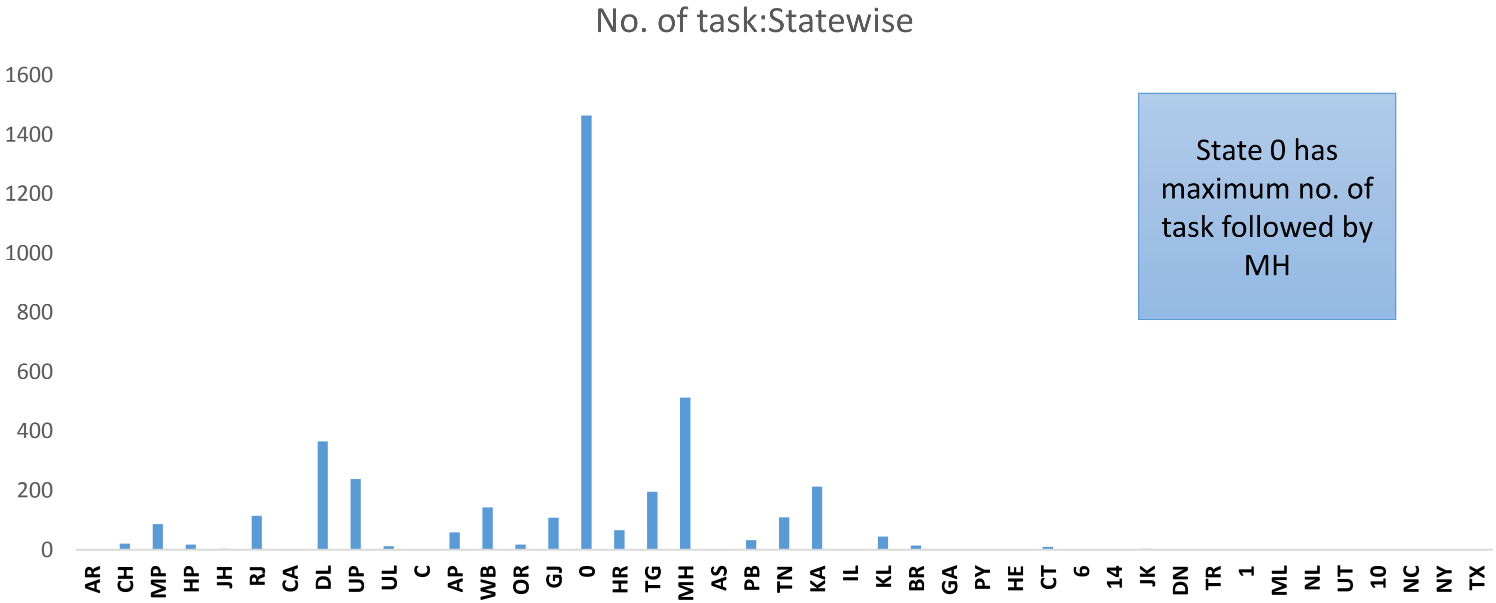
Various plots for exploring the data: No. of tasks



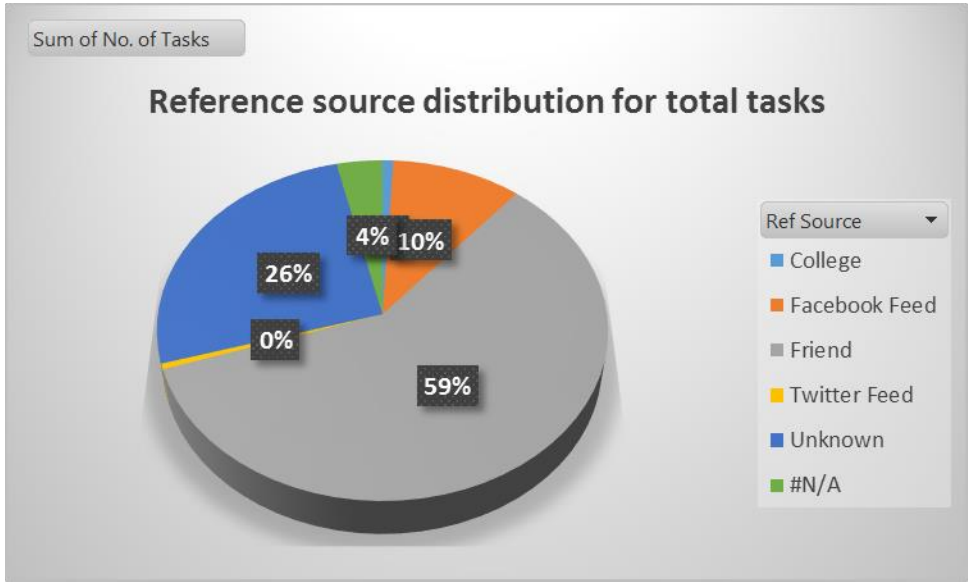
```
 No. of Tasks
Min.   :   1.0
1st Qu.:   2.0
Median :   5.0
Mean   : 104.6
3rd Qu.:  52.0
Max.   :5592.0
```



Most vendors get around 50 tasks in the month



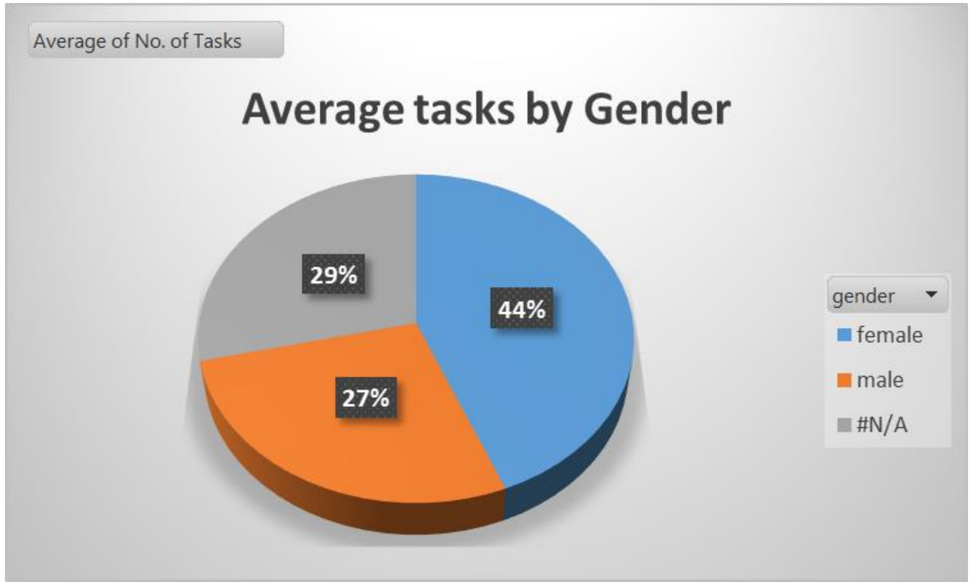
State 0 has maximum no. of task followed by MH



Analysis: No. of tasks

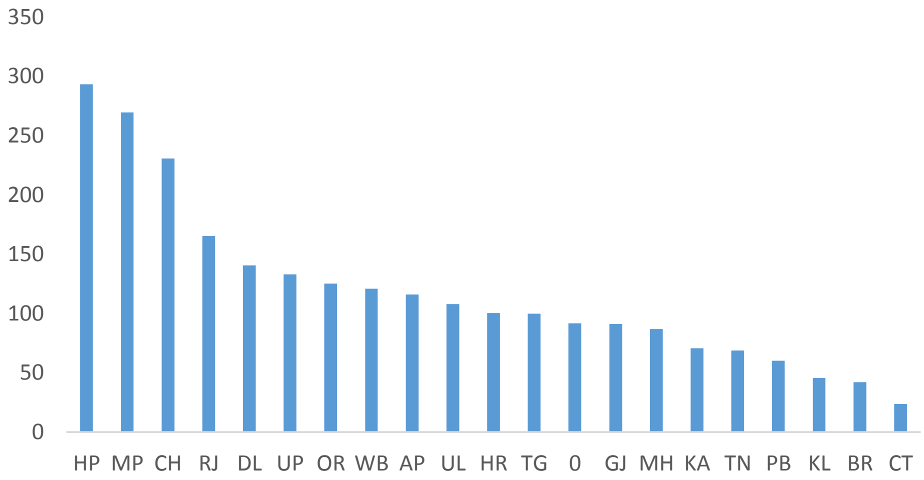
Top 1% user in no. of tasks

User Id	No. of tasks
47554	5592
35911	5548
41968	5081
50188	4170
39791	3904
37862	2814
41150	2792
46048	2656
48422	2582
37741	2580
47057	2553
29675	2274
30161	2083
56153	2058
19547	1945
44666	1904
47811	1880
54143	1844
12938	1827
54944	1819
39228	1818
41418	1814
53154	1767
27492	1751
30635	1722
42654	1709
49649	1639
53758	1608
44521	1607
28335	1607
29412	1590
53123	1586
50245	1584
53880	1583
52745	1547
37771	1529
48183	1463
56321	1451
32996	1434
51669	1429
57334	1429


Average tasks by Gender

Females attempt more no. of tasks compared to males on average

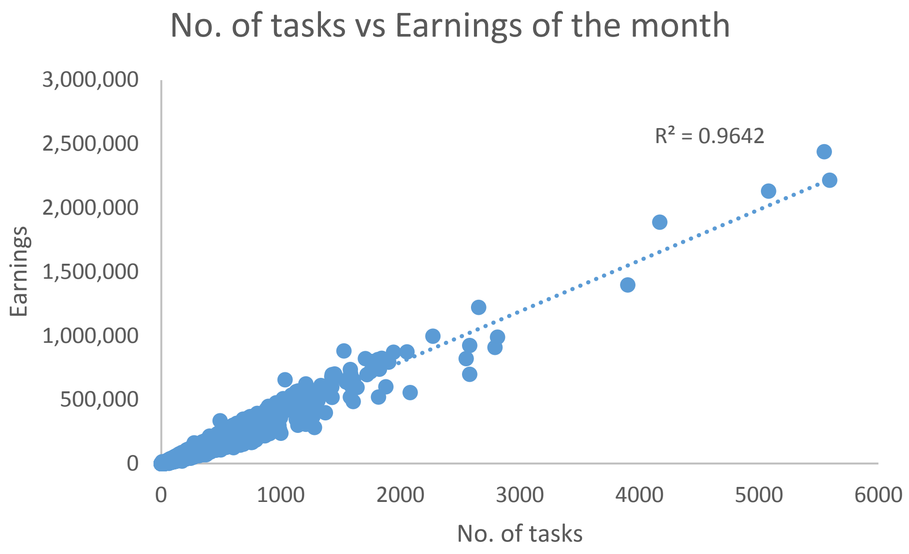
Average of No. of Tasks: By state



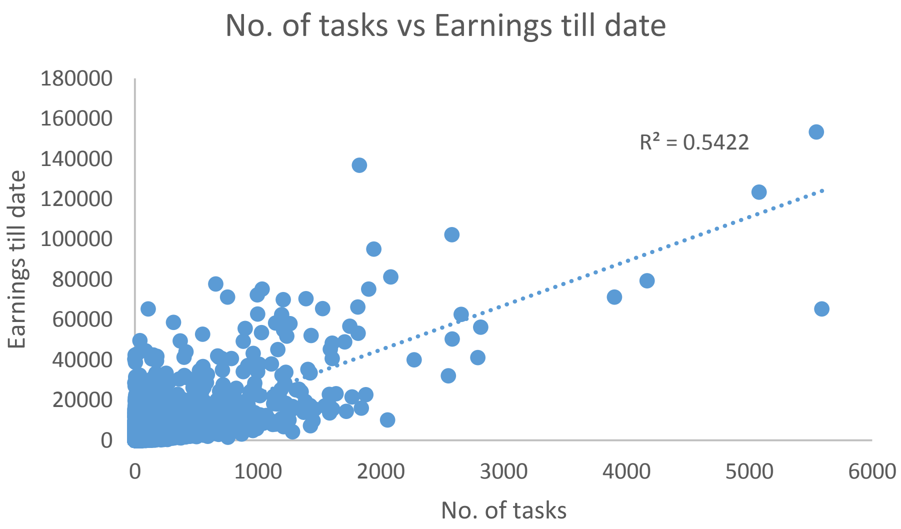
Note: Removed less then 10 tasks to get a correct picture

HP has on average maximum no of tasks attempted across states

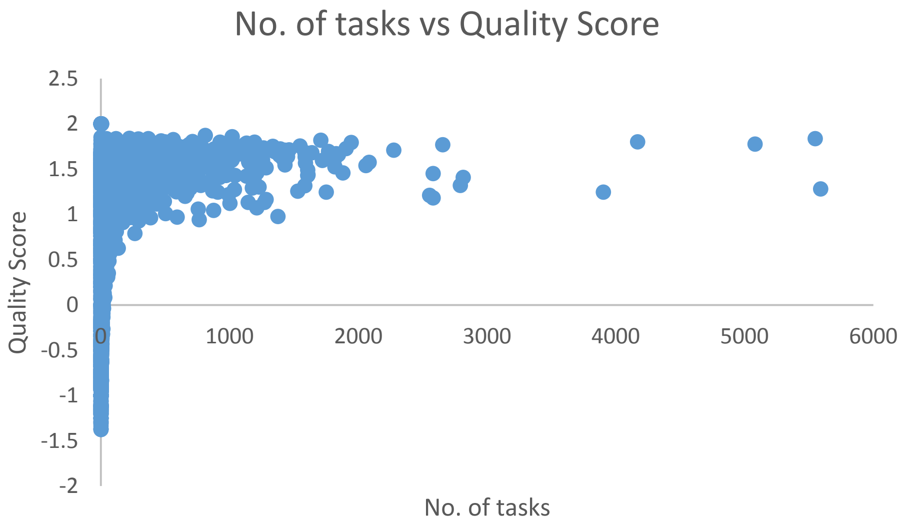
Correlations

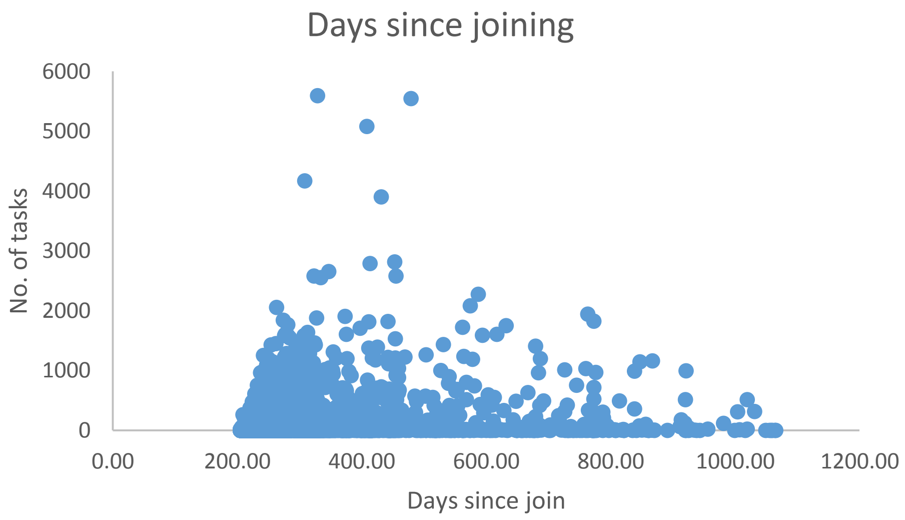

No. of tasks vs Earnings of the month

Strong Co-relation


No. of tasks vs Earnings till date

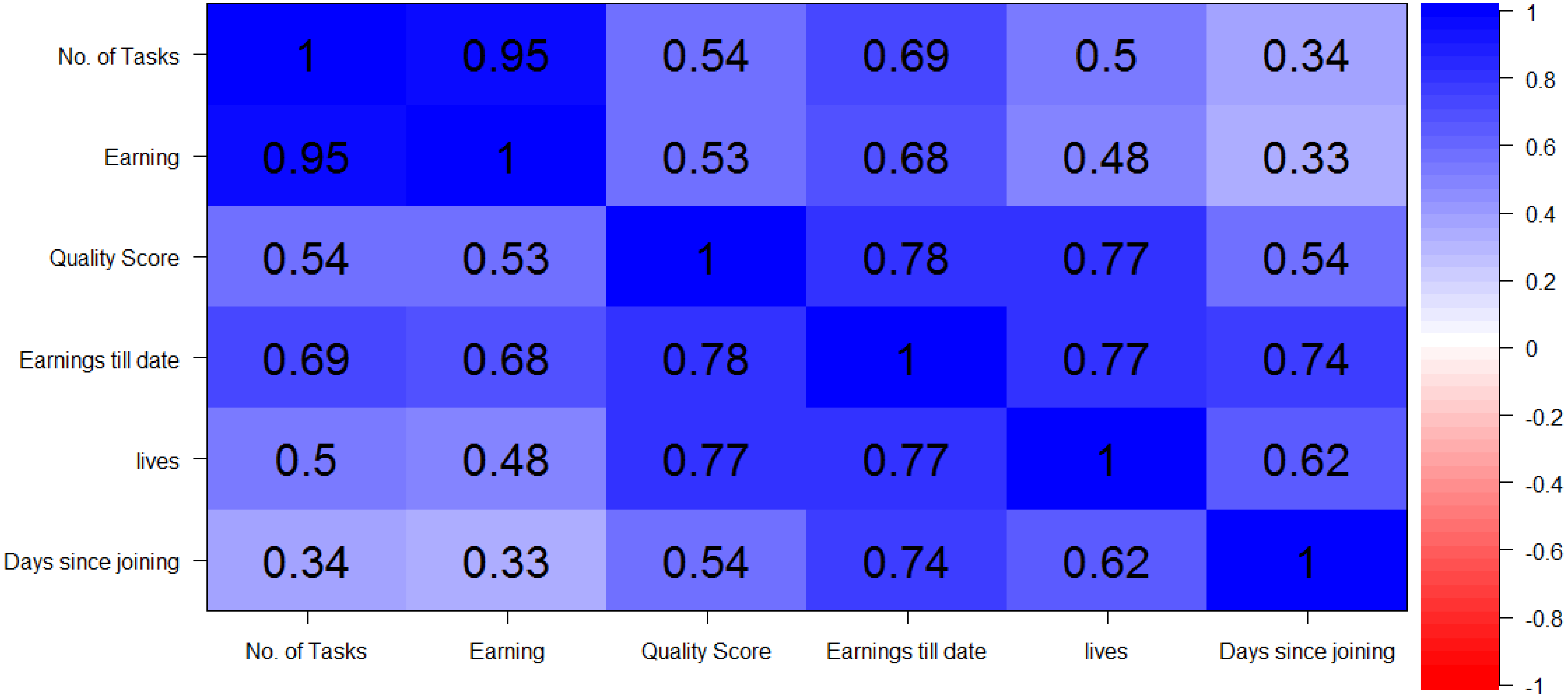
Slight co-relation


No. of tasks vs Quality Score


Days since joining

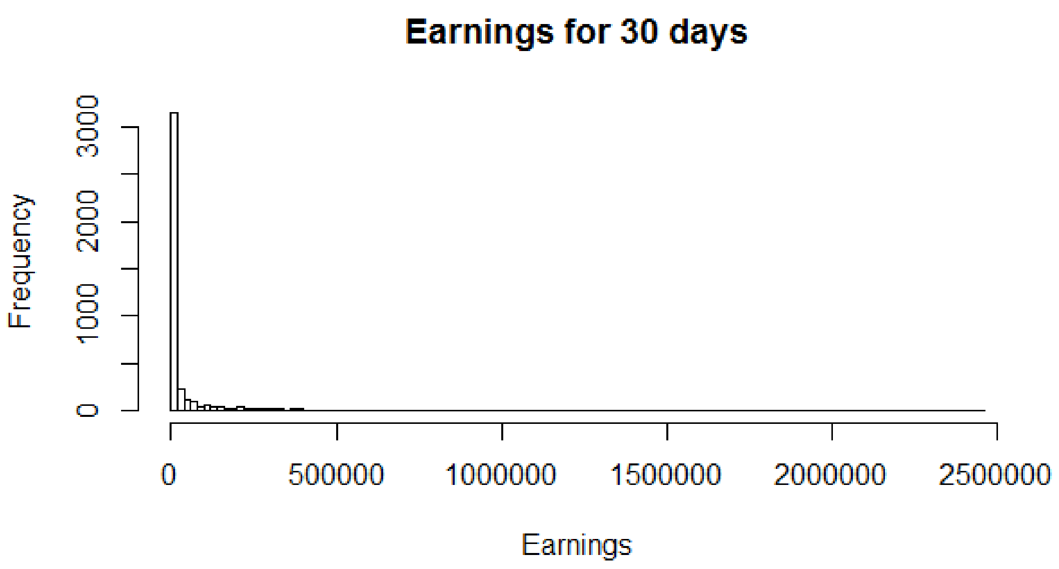
Initial gap of 200 days is may be due to error in calculating no. of days.(Assumed 22-05-2017 as the last date

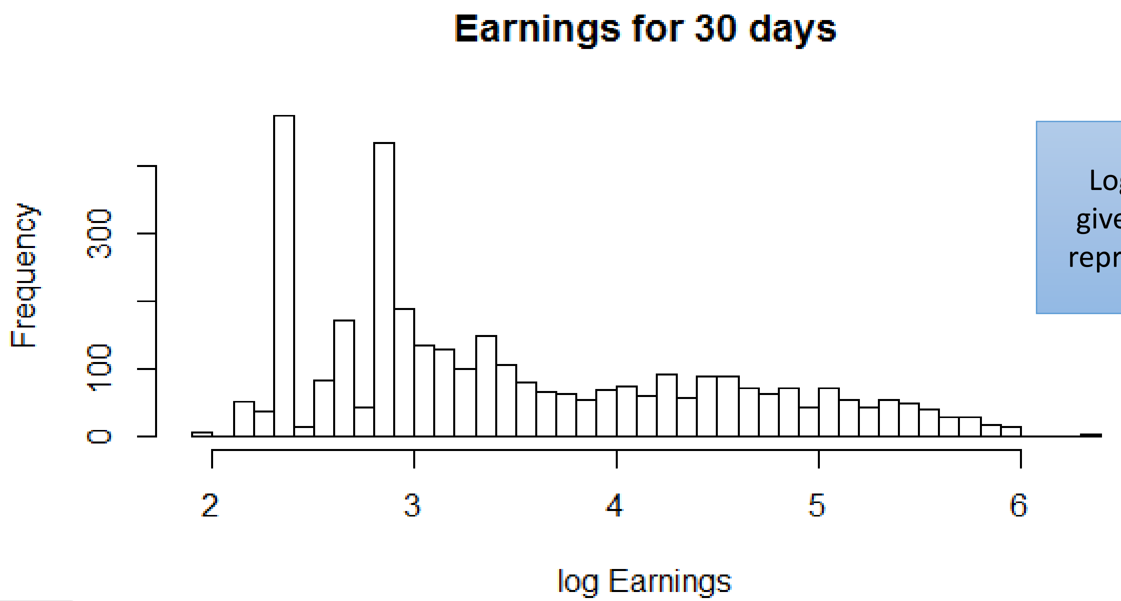
Co-relation matrix among numerical data



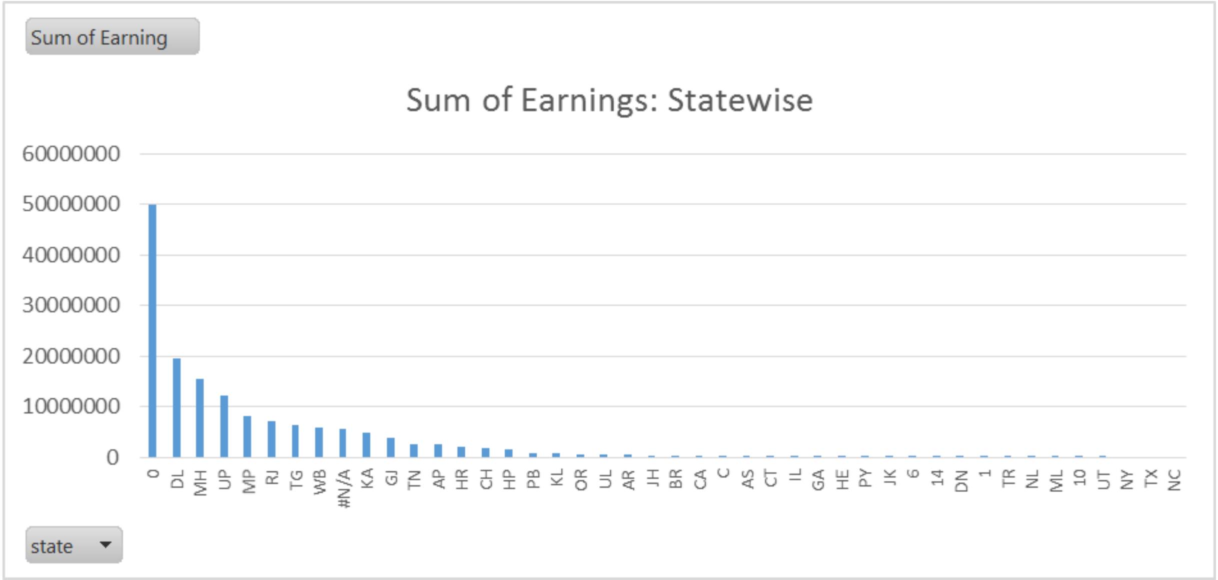
	No. of Tasks	Earning	Quality Score	Earnings till date	lives	Days since joining
No. of Tasks	1	0.95	0.54	0.69	0.5	0.34
Earning	0.95	1	0.53	0.68	0.48	0.33
Quality Score	0.54	0.53	1	0.78	0.77	0.54
Earnings till date	0.69	0.68	0.78	1	0.77	0.74
lives	0.5	0.48	0.77	0.77	1	0.62
Days since joining	0.34	0.33	0.54	0.74	0.62	1

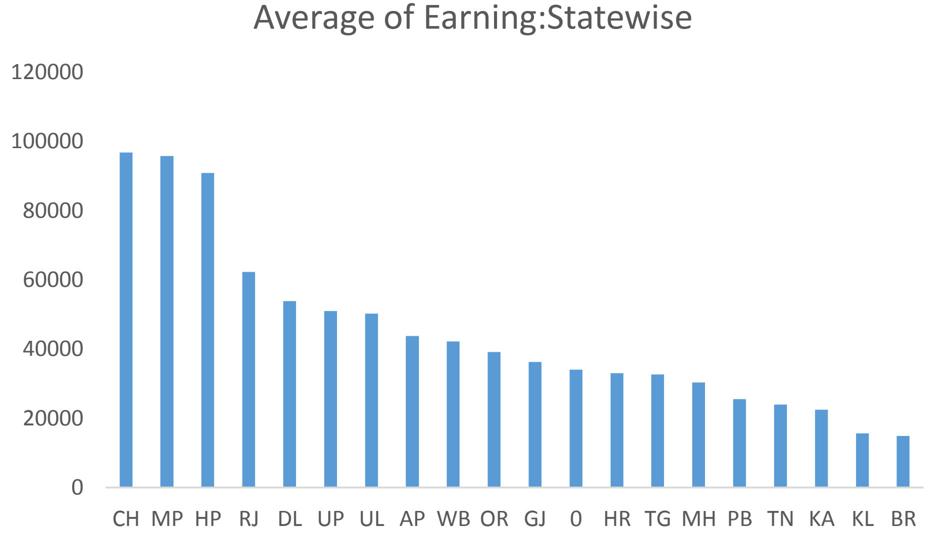
Analysis of Earning: For 30 day period



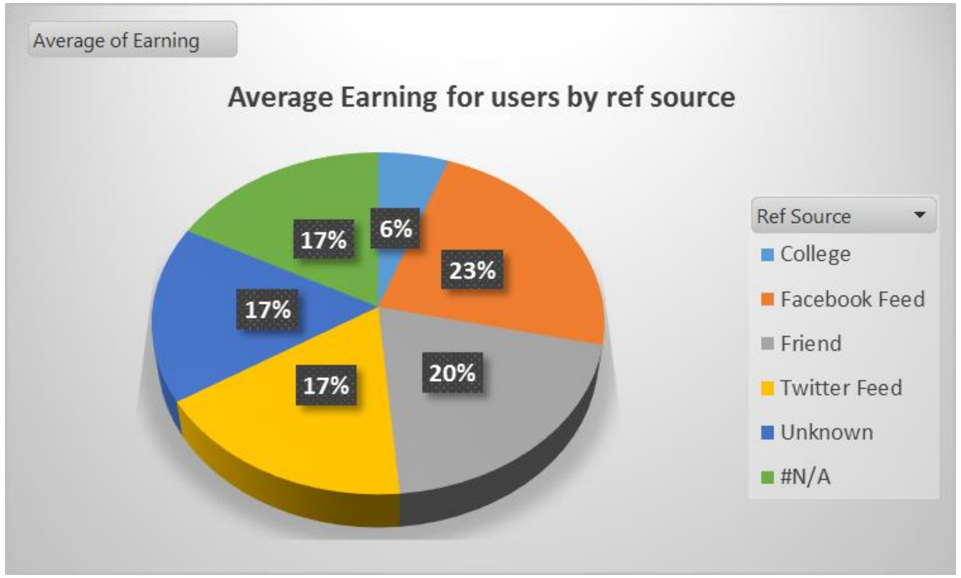


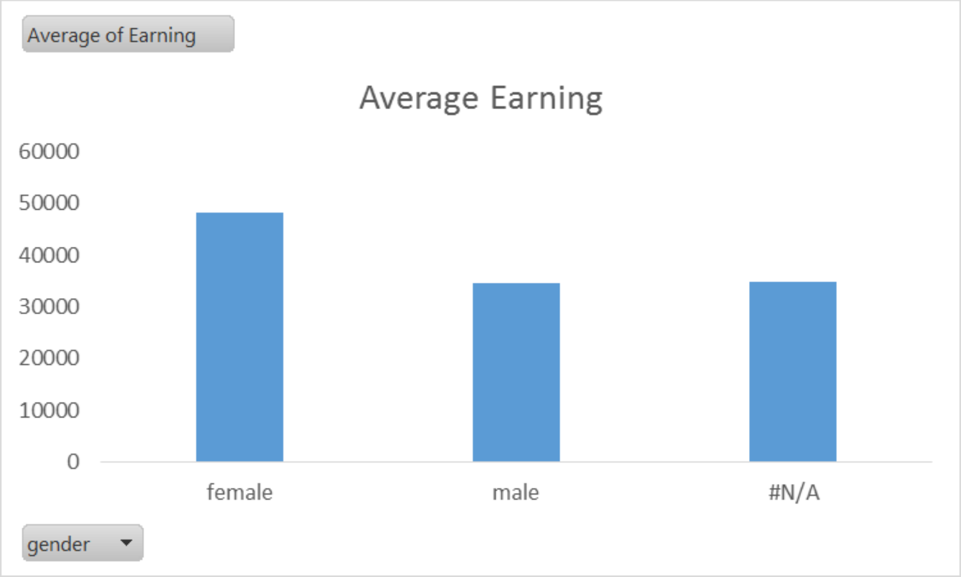
Log Earning gives a better representation





Analysis of Earning: For 30 day period

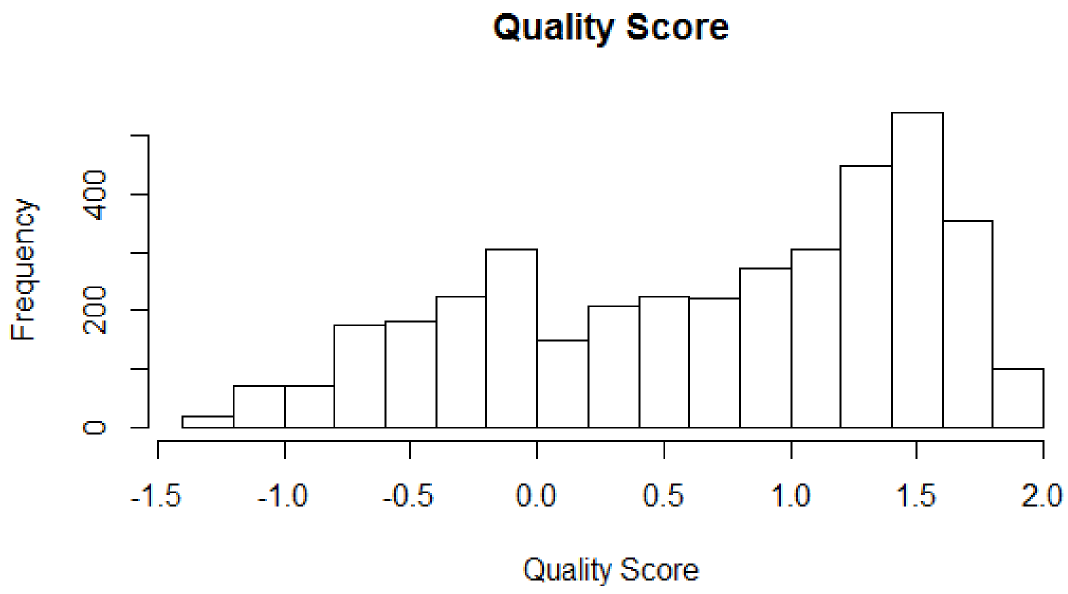


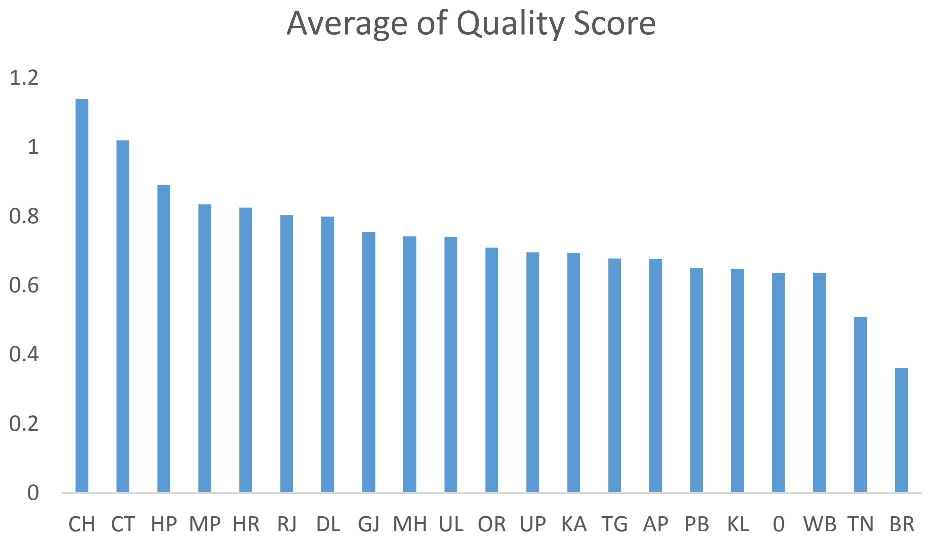


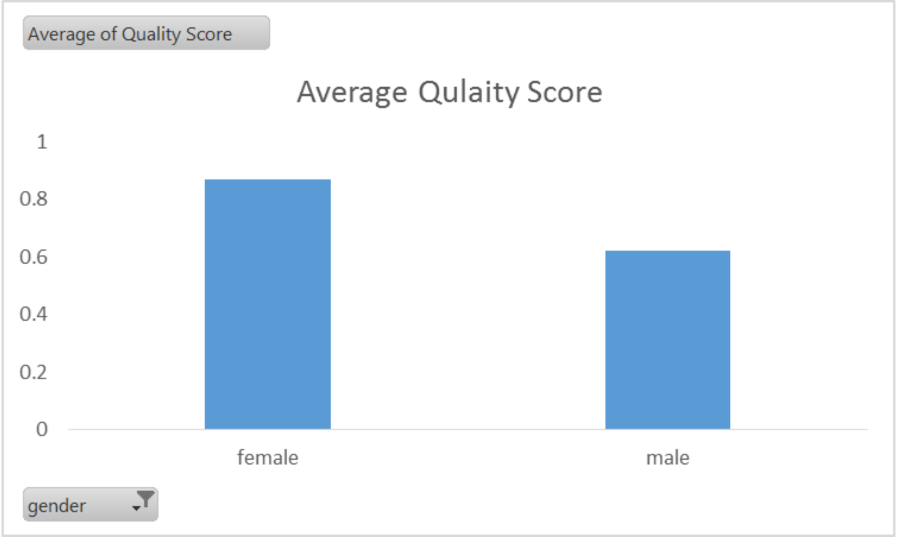
Top 1%

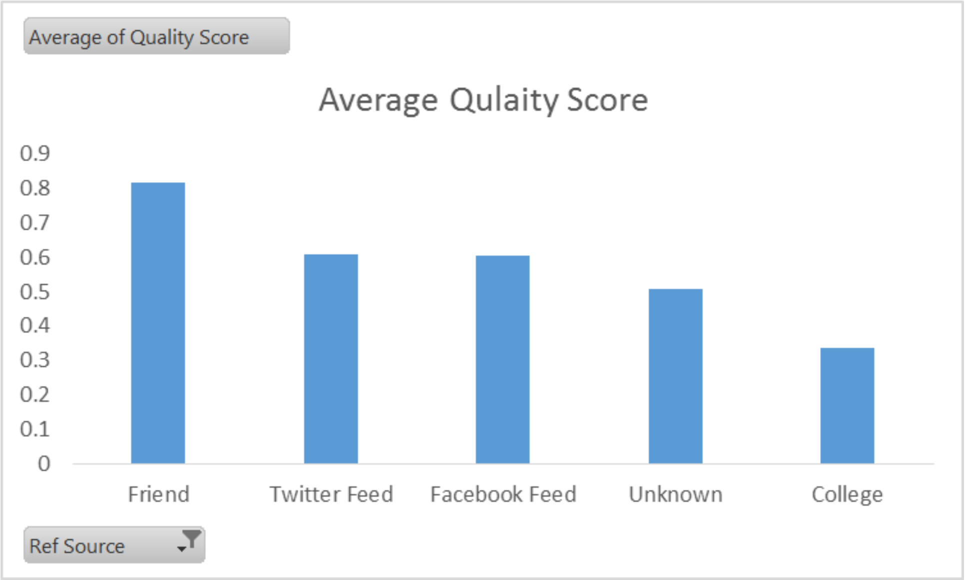
Id	Sum of Earning
35911	2441410
47554	2218538
41968	2131930
50188	1889825
39791	1399290
46048	1223085
29675	998725
37862	990405
37741	925515
41150	911210
37771	881305
56153	874500
19547	871530
54143	824675
42654	821950
47057	821180
41418	815738
53154	800388
44666	795020
54944	754735
12938	739760
53880	738158
27492	719415
53123	701990
56321	700900
48422	700625
51669	697313
30635	697125
32996	687970
51091	682795
48183	679408
53758	657585
37389	656105
52745	638445
28335	634450
37744	624460
49993	611777
29412	608455
40174	605510
51751	601255
47811	601070

Analysis: Quality Ratings









Summary Slide

Parameter	Value
Total No. of Id	4032
Total no. of tasks	421849
Average tasks per id	104.62
Id with max task	Id:47554 Tasks:5592
State with max task	State: 0 Sum of tasks=134230
Total male	2786
Total Female	1084
Average Earning	3439
State with max avg earning	CH
Average Quality Index	.692
State with max avg quality index	CH

- Most vendors get around 50 tasks in the month.
- Females are better in terms of Quality score, average tasks and earnings compared to males.
- State CH tops in terms of average Quality score and average earnings.
- State 0 has maximum no. of tasks.
- User Id referred by friends performs slight better compared to others in terms of no. of task and average Quality score.
- Quality score distribution is approximately bi-modal.
- Higher Quality score correlated with higher lives
- Top 1% user list has been shared. This can come handy while assigning new tasks

Task 2

- Objective:Activity 2:This data set has results from Tests we ran on the platform. There are also results from various missions which have been running on the platform. Based on available data, try to arrive at a framework and/or obtain insights into the performance of a 'Skilled Contractor' and an 'Unskilled Contractor'.

- Elaborate on any one particular metric/characteristic that you would be of particular benefit to us in increasing contractor productivity. Explain why you would choose this particular performance/characteristic/metric.

- This is an open ended activity set and you are free to take any approach that makes sense to you. Please take care to explore and explain the methodology you adopt in detail.

Analysis the data and Design of Metric

Available Parameter

Parameter	**Significance**
p_id	Id of contractor
Gender	
age	
vintage (days)	Duration of association
Status of various tests	logical ability, Reading comprehension, general awareness, attention to detail, pattern recognition to determine aptitude of vendor
Life Time Earnings in Rs.	Total Earnings
rejects	No.of rejects for various tasks
approved	No.of approves for various tasks
re_attempts	No.of re-attempts for various tasks
lives	Lives left

Metric Design:

Current data sheet contains multiple parameters to access a vendor's performance. We can design a metric which includes all these parameters and assigns a final rating to the vendor so that we can segregate various vendors based on this index value.

1) Q Scores: Q scores have been calculated for various task parameter data(using accept, reject, re-attempt) for various tasks to evaluate the accuracy of the vendor as per following formula.

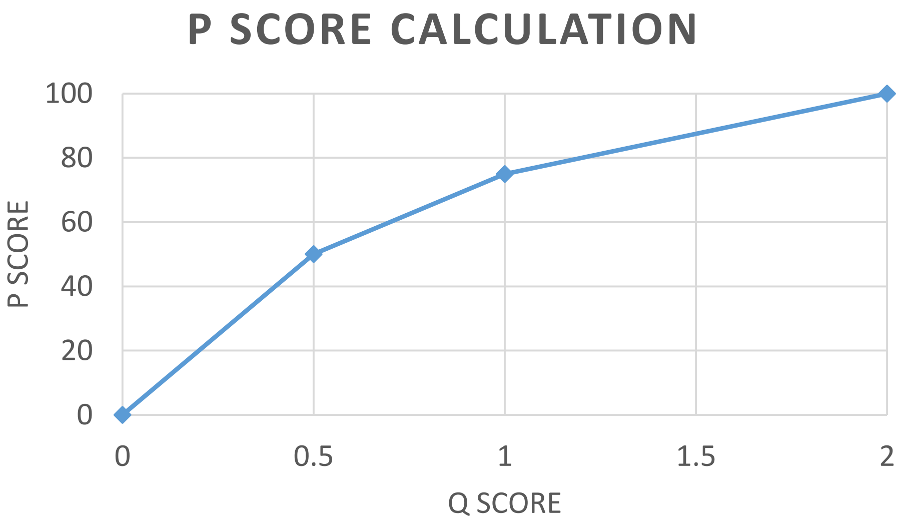
Q_Score=(2*Approved-2*rejects-1*re-attempts) /(Approved+rejects+re-attempts)

The above index gives a value in between +2 to -2 based on the accuracy of the vendor. All approved tasks will result in a score of 2 while all rejects will result in -2. Above concept is taken from the task description for ease of understanding, although more complex index can be constructed as per requirement.

Hence Q score has been calculated for all 5 task which has rejects, approved, re attempt data

P score calculation and final metric

After calculation of Q scores we can take a average of all Q scores to get the final index. However, in doing so we will neglect the consistency/precision of the vendor which is of great importance of measurement.   For example: if the Q score for a vendor is 100 in one task and 20 in other, this will reject in a average index of 60. However the person performed poorly in one of the task but it got averaged out and hence did not come into light. A better matrix is which penalizes non linearly across the Q score range so that we can ensure a better consistency and higher accuracy of the vendor. I have used a exemplary non linear penalization mechanism for the vendor, a more complex system can be designed based on requirement.



So, in the graph we can se that we have different levels of penalty different score of Q no. If the Q no. is between 1 to 2, we have a slope of 1,and then increase 2 fold in subsequent intervals. This ensures that a lower Q score is penalized heavily compared to higher Q score. For a Q score below 0, 0 P score is assigned.
I have also assigned P scores to various aptitude test results by assigning 100 if qualified and 0 if not. So a total of 14 new index's have been calculated consisting of 5 Q scores and 9 P scores. A average of 9 P scores gives the final index for each vendor.

Before averaging we can devise even a more complex metric which penalizes Q scores exponentially as per following equation, **but have not been used**:

P score=10^4*(100-Q Score)

And finally taking anti log of average P score to calculated final index.

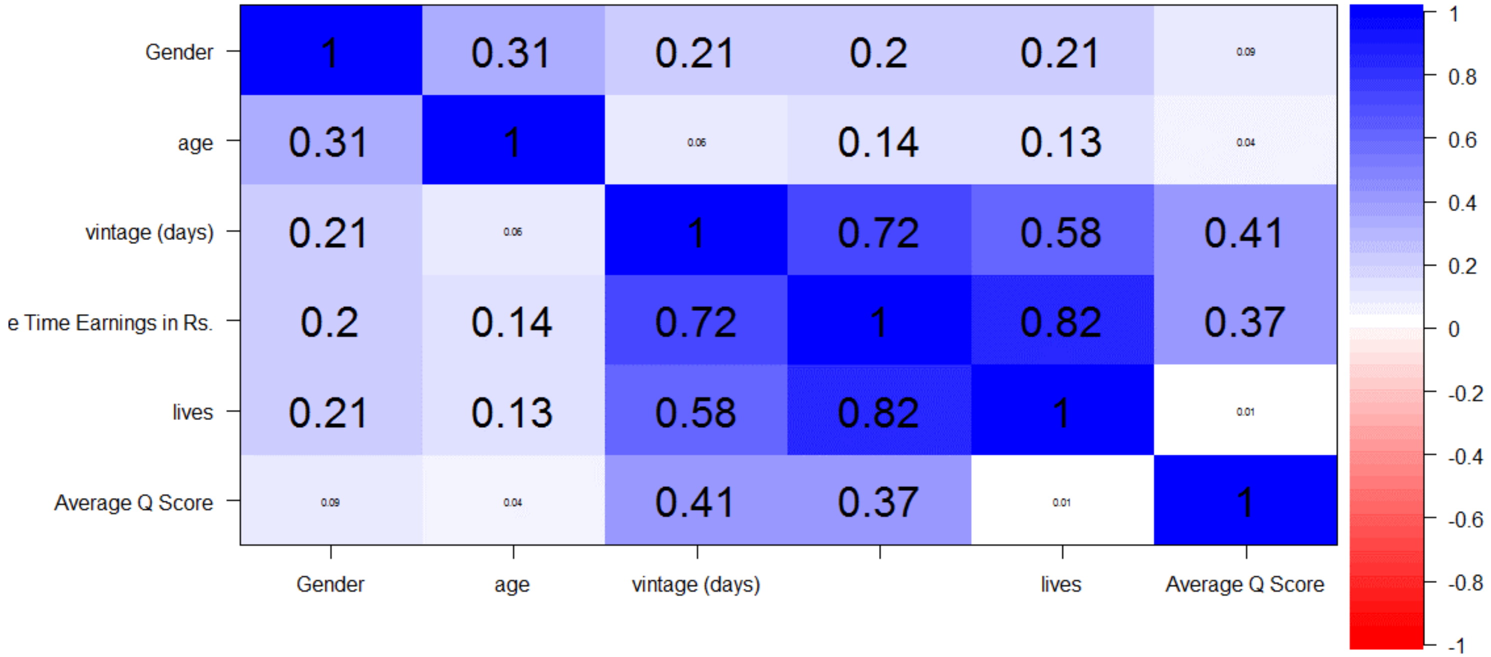
Analysis of metric scores

A excel sheet has been shared which shows the calculated index data for all vendors. Vendors can be segregated as Skilled or Unskilled by using a cut off average index value. A snapshot of average P score for all index's is as follows

Index	Value
Average of P score other data	89.16487
Average of P Score Qualifiers	24.73539
Average of P Score Sd	89.15507
Average of P Score ts	95.65191
Average of P_logical_ability	56.9378
Average of P_reading_comprehension	56.45933
Average of P_general_awareness	63.63636
Average of P_attention_to_detail	68.89952
Average of P_pattern_recog	75.11962
Average of P Score voice	41.39845

Average P score shows that P score Qualifiers and P score Voice are below 50 and can be used as focus index for improvement

Correlation metric



	Gender	age	vintage (days)		lives	Average Q Score
Gender	1	0.31	0.21	0.2	0.21	0.09
age	0.31	1	0.06	0.14	0.13	0.04
vintage (days)	0.21	0.06	1	0.72	0.58	0.41
e Time Earnings in Rs.	0.2	0.14	0.72	1	0.82	0.37
lives	0.21	0.13	0.58	0.82	1	0.01
Average Q Score	0.09	0.04	0.41	0.37	0.01	1

Above metric shows that Average Q Score is slightly correlated to vintage and Earnings, which means a longer duration on the platform improves accuracy. Co-relation metric for P score Qualifiers and P score voice does not yield any significant result.

Predictive analytics

Objective1: Given all the parameters mentioned in slide one we want to predict the Earning for a given Id. This is because in a business context we would like to ensure maximum execution of tasks with highest quality. Since earning is directly calculated based on these parameters it will be a great parameter to predict. In tern we case use the other parameters to select a right group of users to assign the task.

Objective2: Based on task 2 data attempts can be made to develop a predictive model which considers all the important parameters to finally predict average P score for a vendor.

Status: In progress

Expected completion: By 25-05-2017 EOD