Teradata Data Challenge 2020

Project Report

taproot FOUNDATION

TABLE OF CONTENTS

List of Figures

Executive Summary

The Taproot Foundation drives social change by leading, mobilizing, and engaging professionals in pro bono service. Essentially, they are helping nonprofits and social change organizations solve critical challenges in their communities with the support of skilled volunteers sharing their expertise pro bono. Nonprofits and social change organizations take on critical issues facing our communities every single day like homelessness, economic inequality, violence, civil rights, and more. However, they have limited access to the required resources such as marketing, technology, human resources, or funding (Taproot, 2020). Taproot acts as connectors to bring together these organizations with business professionals who volunteer their expertise pro bono. Through this unique collaboration, nonprofit professionals and skilled volunteers work together to build a strong infrastructure that supports the organization in achieving its mission. Up to now, Taproot has served 7,649 organizations and made 13,256 total number of engagements, delivering 1.7 million hours of services with a value of $204 million (Taproot, 2020).

With this service nature, it is very important for the Taproot Foundation to evaluate the effectiveness of their pro bono programs currently offered to nonprofit organizations. The Foundation wants to better understand if their pro bono services (Taproot Plus projects, Taproot Plus sessions) in Business, HR, Marketing, and IT categories can significantly support nonprofits in maintaining and developing critical operating functions. With regard to this, our study aims to explore the project and session data to determine the project types that produce the best outcomes and also find out the number of sessions that nonprofits create before the start of a project. Additionally, the study examines the time to complete each type of project, as well as the average days to complete projects with multiple volunteers. Our exploratory analysis shows that Business and HR are producing better outcomes as compared to IT and Marketing categories. While Evaluation is the best performer in the Business category, Staff and

Board Development projects are the leaders in the HR category. In the IT field, CRM is the best outcome producer and Messaging projects are outperforming others in the Marketing category. Considering the project type across all categories, the CRM projects (which belong to IT) can be considered the best performer overall. Regarding the sessions conducted, only about 0.3% of the non-profits create at most one session before the start of a project. Meanwhile, most of the non-profits do not conduct any session before project creation. Among all project types, Public Relations projects take the longest time to complete (257 days on average) while Design projects take the shortest time to complete (152 days on average). The average number of days taken to complete projects with multiple volunteers is approximately 247 days.

The study's analysis findings expect to provide useful insights for the Taproot Foundation to better understand the effectiveness of their pro bono services with Taproot Plus projects and Taproot Plus sessions and to ensure that available resources are directed efficiently. Besides, identifying underperforming projects can also help the Foundation investigate the cause of the problem and determine if any actionable plan needs to be done in order to fix the issue and improve the project outcomes. This will help the Taproot Foundation further envision all organizations with promising solutions, enabling them to successfully take on urgent social challenges and deliver on their missions more efficiently and effectively.

Business statements

Our analysis focuses on the four business questions as below:

- Which project types produce the best outcomes in survey data?
- How many sessions does a non-profit create before starting a project?
- Which project types take the longest to complete? Shortest?
- Average days to complete projects with multiple volunteers?

Data Description

1. Data access

Our data for this project are exclusively provided by the Taproot Foundation and can be accessed from Teradata 2020 Data Challenge website. There are multiple data sets that provide a variety of Taproot Plus projects and session information as following.

a. User activity data

- *Analytics all Website data pages 2014-2019* (Pageview, Unique Pageview, Avg. Time on Page, Bounce rate, % Exit, Page Value)
- *Conversions* (ID, Service, Channel, Campaign, Context, Terms, User_ID, Convertible_Type, Convertible_ID, …)

b. Session data

- *Session export* (ID, State, Description, Consultant_id, Nonprofit_id, Scheduled_for, Organization_id, Conference_line_id, Project_category_id, Archived, …)

c. Project data

- *Project categories* (ID, Group_slug, Enabled, International, Name, Slug)
- *Project inquiries* (ID, User_ID, Project_ID, Qualifications, State, Created_at, Updated_at, Scheduled_for, Decision_deadline, …)

- *Project export* (ID, User_ID, Organization_Id, Description, State, Campaign_id, Project_inquiries_count, Project_category_id, Local_only, Satisfaction_rating, …)

d. Project recommendation data

- *Project recommendation* (ID, User_ID, Project_ID, Created_at, Updated_at, Context)

e. Email notification data

- *Email notification* (ID, Kind, Scheduled_for, Enqueued_at, Created_at, Updated_at, Notifiable_type, Perform_deliveries, Processed, Delivered, Bounce, Open, User_ID, Message_ID, …)

2. Data preparation

To find the solutions to our business questions, we have combined the session data and project data together. While the project export data provides basic information of each project, the project categories data specifies which category the project falls under and the project inquiries data examines users' engagement on these projects. The session export data also uses the project categories data to know which category the session falls under.

After all data sets are consolidated, some preliminary analyses are performed to identify if there are any missing values/ erroneous data or any potential outliers that need to be taken care of to make sure the master data set is ready for analysis.

Analysis and results

1. Best outcomes by project types

Since there is no definition of '*best outcomes*' in the data dictionary or business questions, and given this term is a subjective measurement, hence, for the purpose of our analysis, we measured '*best outcomes*' as a project being **completed** rather than being rescheduled or postponed.

We first joined the project categories data and the session data together, then we combined several phases of the project to three basic levels as below:

- 'canceled, 'expired' and 'missed' were categorized as '*Not Completed*'
- 'applied', 'draft', 'rescheduled', 'pending' were categorized as '*In Progress*'
- 'published' and 'completed' were categorized as '*Completed*'

Based on the three derived levels, we examined the frequency of each level in a project's life for different project categories (Business/ HR/ IT/ Marketing). The results are provided below.

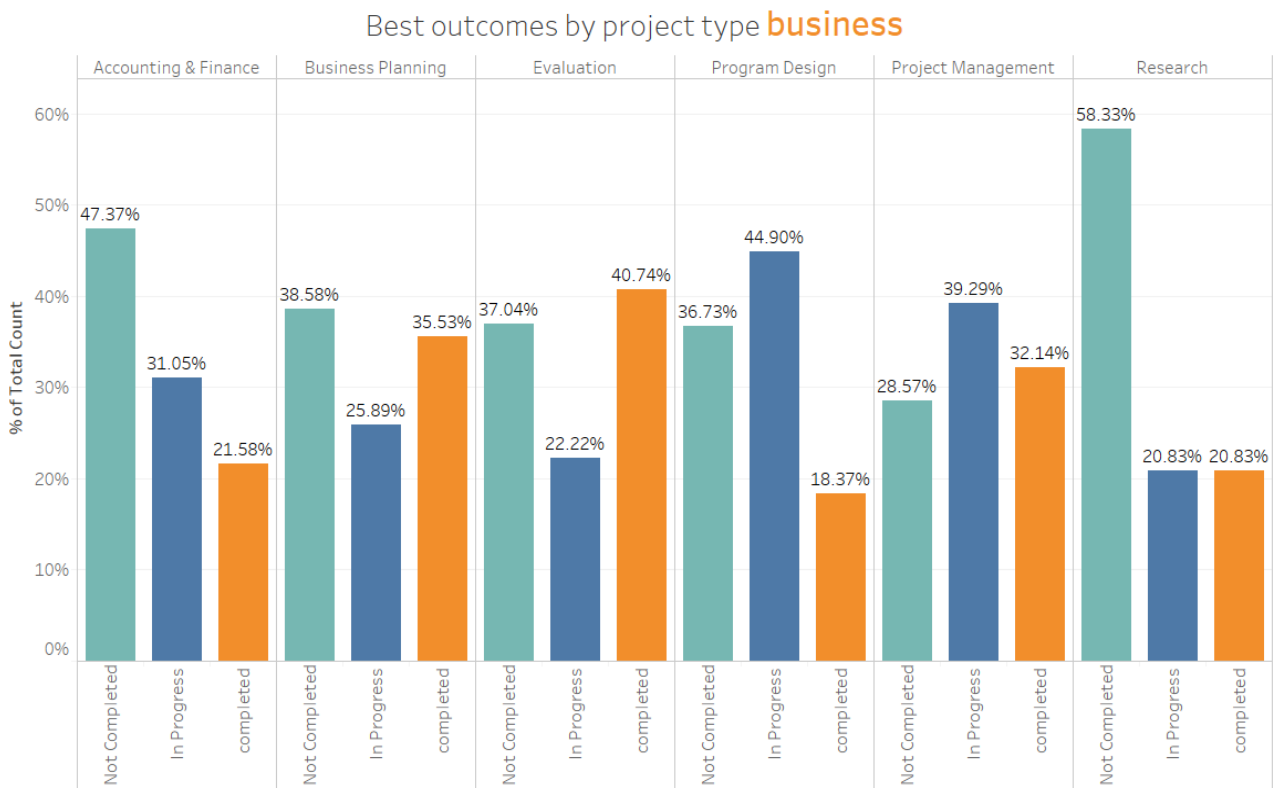


Figure 1 – Best outcome by project type – Business category

For business projects, Business Planning and Evaluation are the two project types that produce much better outcomes compared to other business-type projects. The proportions of three levels for Business Planning and Evaluation are rather similar. However, given our mentioned definition of 'best outcomes', Evaluation projects seem to produce the best outcome with the highest proportion of completed projects at 40.74%, about 5% higher than Business Planning. Evaluation also has lower non-completed projects as compared to Business Planning. Hence, Evaluation projects are considered the best performer in the Business category.

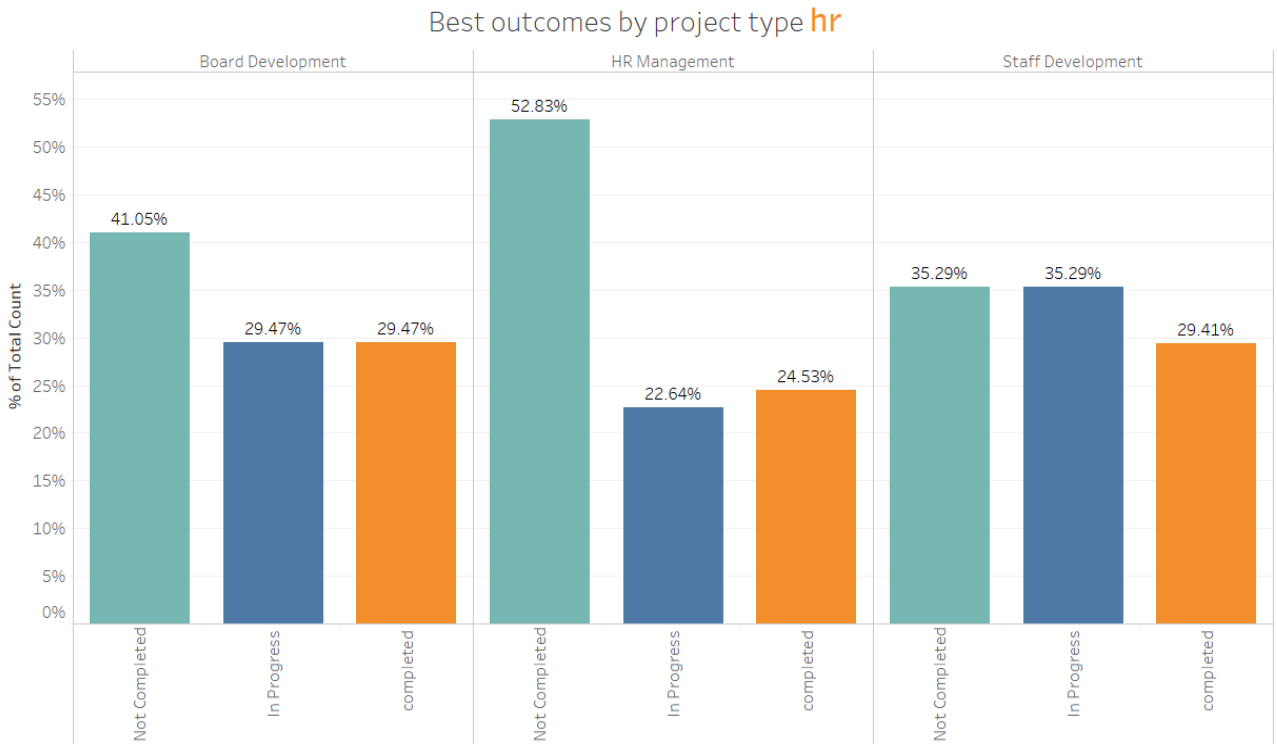


Figure 2 – Best outcome by project type – HR category

For HR projects, Board Development and Staff Development seem to produce the best outcomes with a higher percentage of completed projects (approximately 29.5%) as compared to HR Management.

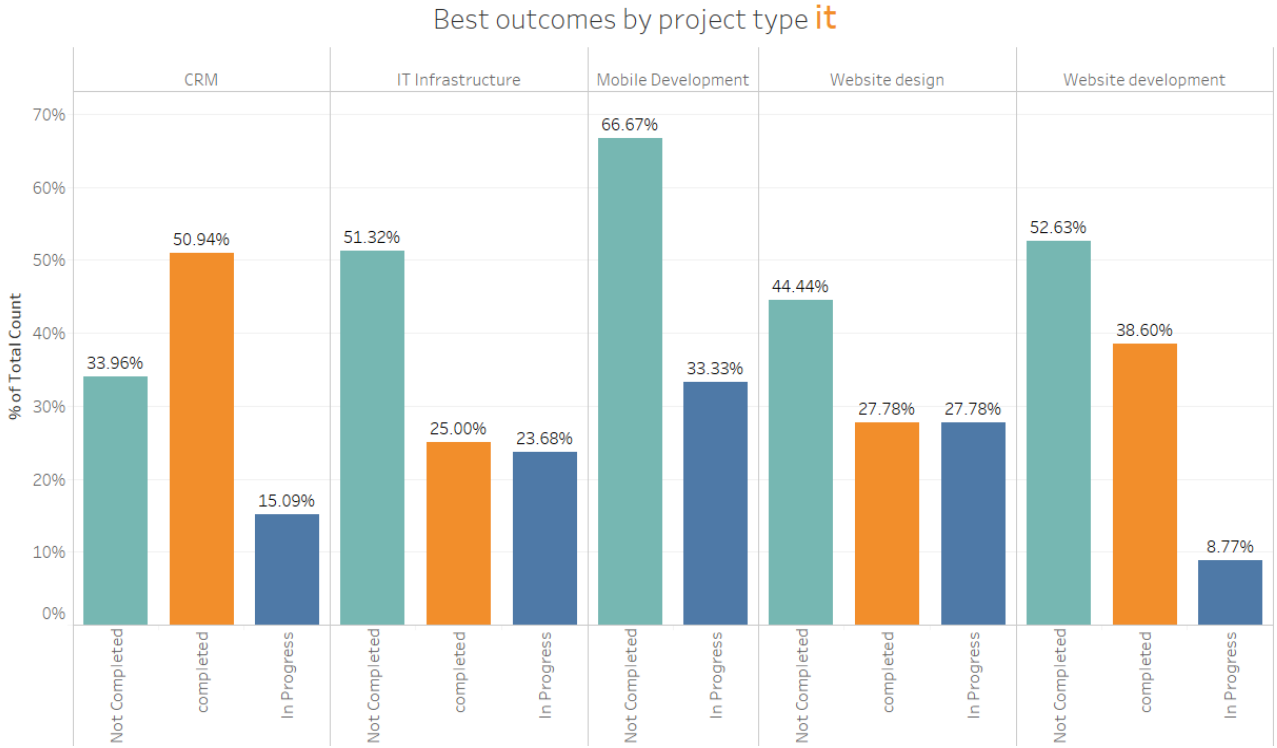


Figure 3 – Best outcome by project type – IT category

For IT projects, CRM is producing the best outcome with 50.94% projects completed as compared to other types. CRM also has the lowest percentage of non-completed projects; thus, it can be considered as the best performer in the IT category.

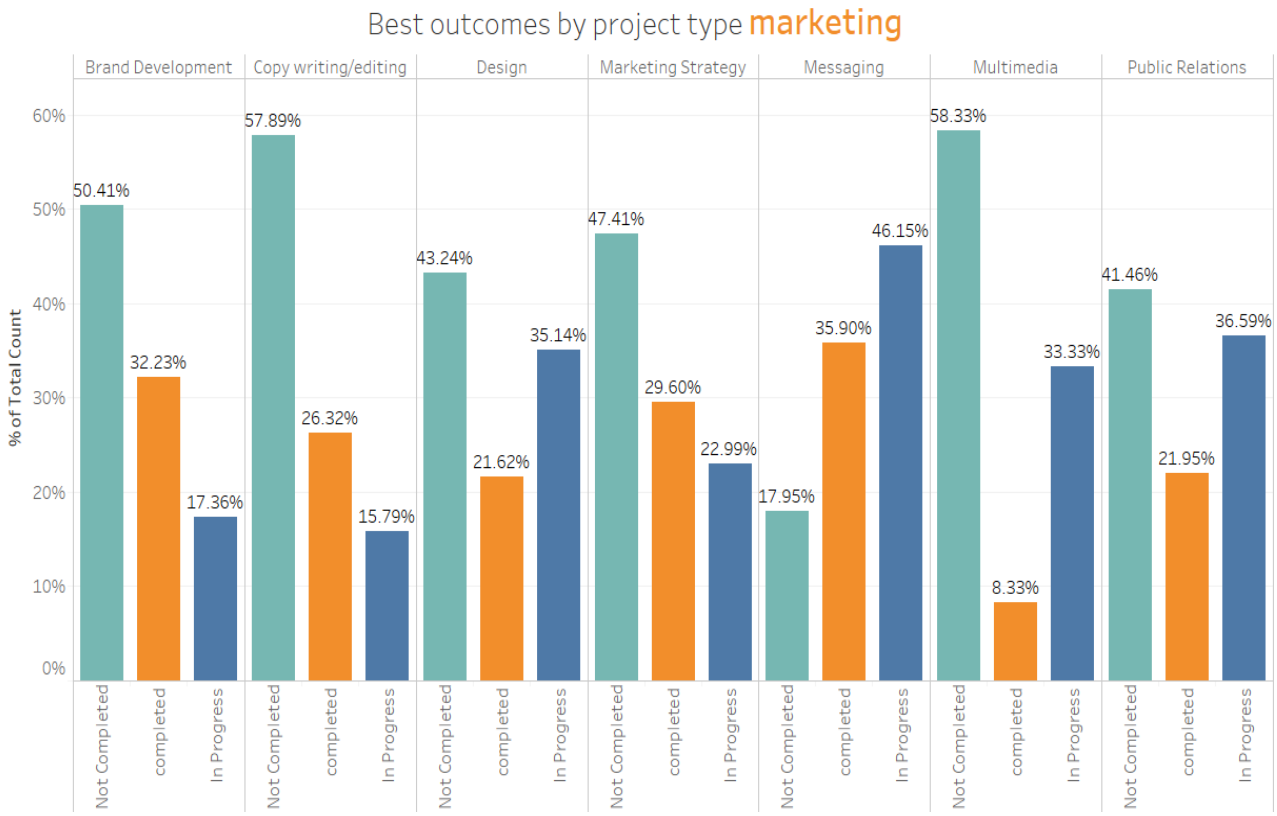


Figure 4 – Best outcome by project type – Marketing category

For Marketing, among seven project types, Messaging is likely the best outcome producer with 35.90% completed projects as compared to other types. Like CRM in IT, Messaging also has the lowest proportion of non-completed projects in Marketing.

Overall, among four project categories (Business/ HR/ IT/ Marketing), Business and HR seem to produce better outcomes as compared to IT and Marketing categories. Considering the project type across all categories, the CRM projects (which belong to IT) can be considered the best performer with the highest percentage of completed projects as compared to other project types.

2. Number of sessions created by non-profits before starting a project

To answer this question, we examined the session export dataset, in which we determined the starting time of each project based on the created date and session date derived from the time slot field in the data. In the next step, we calculated the number of sessions created by each nonprofit before the computed starting date of each project. The result is provided in Figure 5 below.

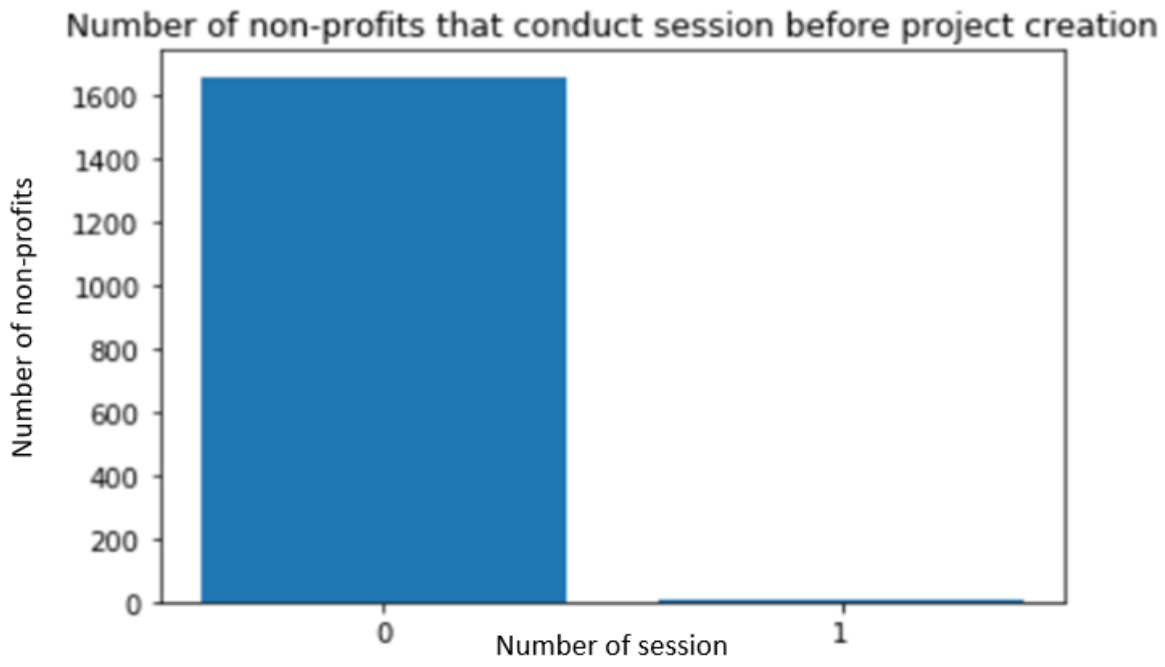


Figure 5 – Number of non-profits that conduct sessions before starting a project

Error! Reference source not found. Figure 5 indicates that only about 0.3% of the non-profits conduct at most one session before starting the project. Meanwhile, most of the non-profits do not create any sessions before the start of the project.

3. Average time to complete a project by project type

We examined the length of each project type by calculating the difference between the created time and the updated time of all completed projects. The average days to complete each project type are provided in the below output.

```
name
Design                    151.701613
IT Infrastructure         156.800000
Accounting & Finance      158.331754
Staff Development         167.508772
HR Management             167.695946
Copy writing/editing      179.625000
Evaluation                183.629630
CRM                       188.137795
Brand Development         201.912791
Board Development         203.620000
Website development       205.009615
Messaging                 207.539568
Research                  208.346535
Website design            212.767528
Program Design            216.154762
Business Planning         219.415550
Multimedia                222.951613
Marketing Strategy        227.346789
Project Management        239.133333
Mobile Development        254.344828
Public Relations          256.896552
Name: Difference, dtype: float64
```

Figure 6 – Time to complete the project

The results show that Public Relations projects take the longest time to complete (approximately 257 days on average) while Design projects take the shortest time to complete (about 152 days on average). Public Relations and Design both belong to Marketing projects.

Figure 7 and Figure 8 below further show the distributions of days taken to complete Design and Public Relations projects.

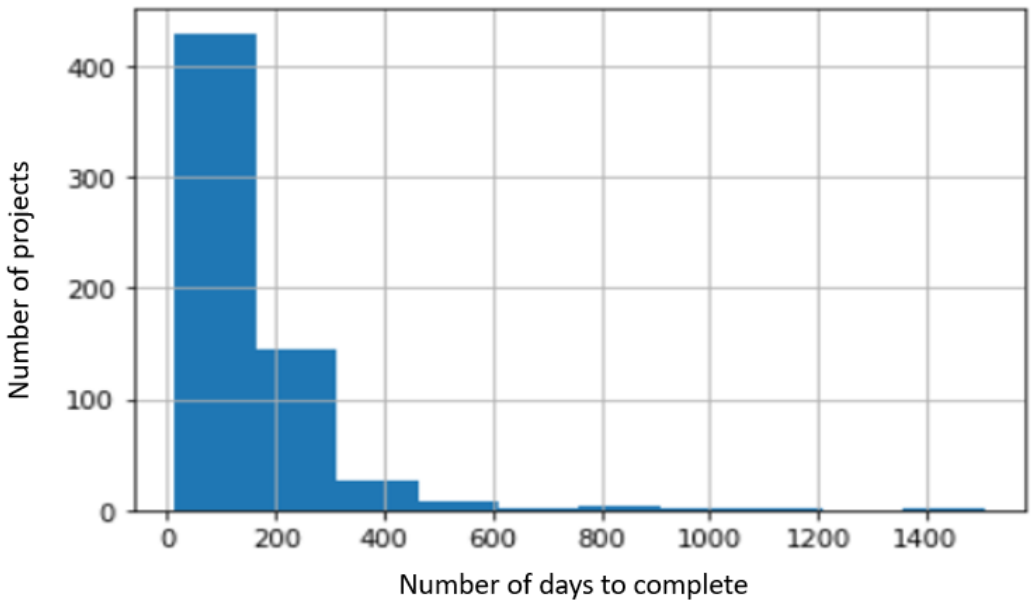


Figure 7 – Distribution of days taken to complete Design projects

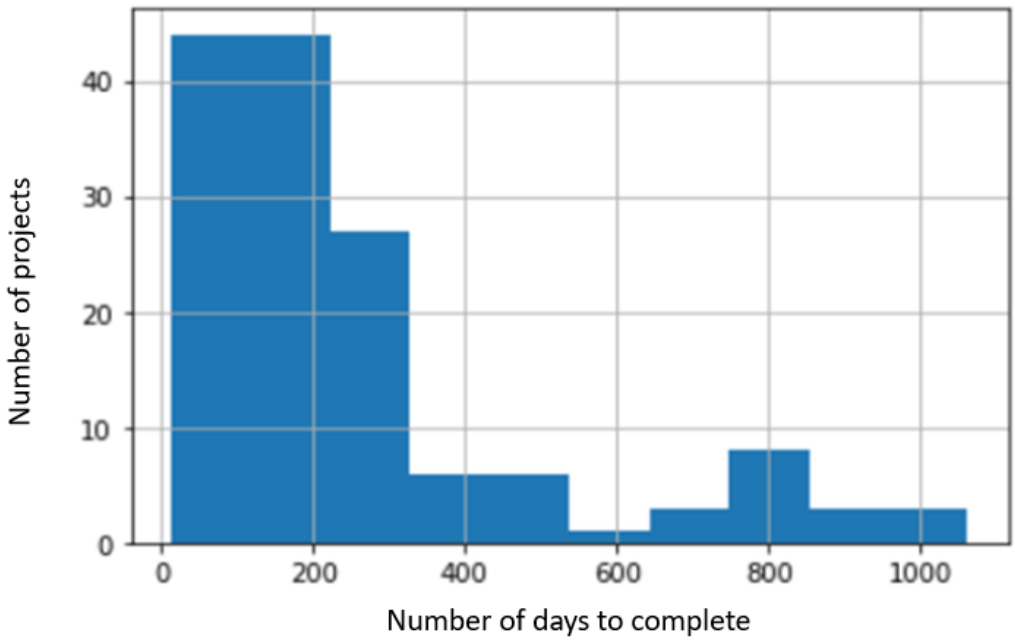


Figure 8 - Distribution of days taken to complete Public Relations projects

4. Average time to complete projects with multiple volunteers

To solve this question, we looked into the Project inquiries data in which the projects that have at least two volunteers in the 'applied', 'confirmed', or 'completed' status were selected. This data was then joined with the Project export data, after which the projects which were 'completed' and 'published' were considered. The time to complete a given project was obtained by computing the day difference between the updated date and the start date.

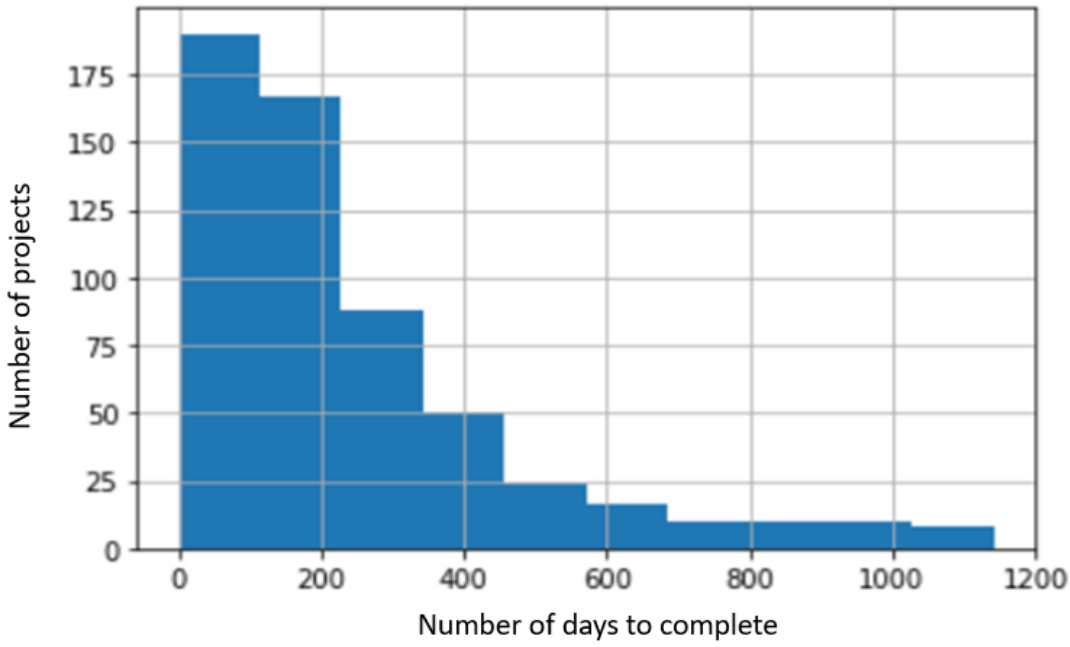


Figure 9 - Distribution of days taken to complete projects with multiple volunteers

Figure 9 above shows the distribution of days taken to complete these projects. As calculated, the average number of days taken to complete projects with multiple volunteers is 246.89 days.

Conclusions and Recommendations

As nonprofit organizations do not always have access to the marketing, HR, technology, or planning resources they need to tackle the issues facing our communities, hence, Taproot Foundation has been helping them to get better access to these resources with the support of skilled business professionals sharing their expertise pro bono. Among the four project categories (Business/ HR/ IT/ Marketing), our analysis shows that Business and HR are producing better outcomes overall. While Evaluation is the best performer in the Business category, Staff and Board Development projects are the leaders in the HR category, and Messaging is the best outcome producer in the Marketing field. Considering a variety of project types across all categories, CRM projects (the best performer in IT) are producing the best outcomes with the highest percentage of completed projects as compared to other project types. Regarding the number of sessions, only 0.3% of nonprofits conduct at most one session before the start of the project. As for project duration, Design projects are taking the shortest time and Public Relations are taking the longest time to complete. On average, it took about 247 days to complete a project with multiple volunteers.

Given these findings, a couple of recommendations are suggested as below.

- Further research needs to be performed for underperforming project types (Research, HR Management, Mobile Development, Website Development, IT Infrastructure, Multimedia, …) in order to identify any issue (for example, more volunteers needed, more funding required, any technical problem, …) and propose possible solutions accordingly
- Increase the number of sessions conducted at nonprofits before the project starts if necessary

- Examine the project types that take a long time to complete (Public Relations, Mobile Development, Project Management, Marketing Strategy, Multimedia, …) and find out any issues that need to be tackled

These recommendations are expected to help Taproot Foundation further strengthen their pro bono program between non-profit organizations, corporates, and volunteers in order to achieve the triple win, in which nonprofits get the support they need, corporates can invest more deeply in their communities, and employees have the professional opportunities they want.

References

Taproot Foundation, 2020. "Nonprofits". Accessed March 2020.

https://taprootfoundation.org/nonprofits/

Taproot Foundation, 2020. "Taproot in numbers". Accessed March 2020.

https://taprootfoundation.org/

Appendix

Question 1

```
# Read data into Python

import pandas as pd

d1 = pd.read_excel("C:\\teradata\\project_categories.xlsx")

d2 = pd.read_excel("C:\\teradata\\session_export.xlsx")

d1.drop(columns = ['created_at', 'updated_at', 'enabled', 'international', 'slug'], axis = 1,
inplace = True)

d1.head()
```

	id	group_slug	name
0	1	business	Accounting & Finance
1	3	business	Evaluation
2	5	business	Program Design
3	6	business	Research
4	7	hr	HR Management

```
d2.drop(columns = ['id', 'created_at', 'updated_at', 'description', 'consultant_id', 'nonprofit_id',
'scheduled_for', 'time_slots', 'organization_id', 'conference_line_id', 'partner_organization_id', '
archived'], axis = 1, inplace = True)

d2.head()
```

	state	project_category_id
0	cancelled	20.0
1	completed	20.0
2	cancelled	20.0
3	completed	10.0
4	completed	20.0

```
#Merge Session data and Project_categories data

d3 = pd.merge(d2, d1, how = 'left', left_on = 'project_category_id', right_on = 'id')

d3.shape
```

d3.head()

	state	project_category_id	id	group_slug	name
0	cancelled	20.0	20.0	it	Website development
1	completed	20.0	20.0	it	Website development
2	cancelled	20.0	20.0	it	Website development
3	completed	10.0	10.0	marketing	Marketing Strategy
4	completed	20.0	20.0	it	Website development

#Group the data by project category, project type, and project status and calculate frequency accordingly

```
d4 = d3.drop(columns = ['project_category_id', 'id'], axis = 1)
```

```
d4.groupby(['group_slug', 'name', 'state']).size()
```

```
d4.groupby(['group_slug', 'name'])['state'].value_counts(normalize = True).sort_values(ascending = False)
```

```
group_slug  name                  state
business    Accounting & Finance  applied              9
                                  cancelled           18
                                  completed           41
                                  draft               26
                                  expired             68
                                  missed               4
                                  npo_rescheduled      3
                                  pending             20
                                  published            1
            Business Planning     applied              8
                                  cancelled           18
                                  completed           70
                                  draft               26
                                  expired             57
                                  missed               1
                                  pbc_rescheduled      4
                                  pending             12
                                  published            1
            Evaluation            applied              1
                                  completed           11
                                  draft                3
                                  expired              9
                                  missed               1
                                  pending              2
            Program Design        applied              3
                                  cancelled            1
                                  completed            9
                                  draft               10
                                  expired             17
                                  matched              2
                                                    ...
```

```
                                                ...
marketing   Marketing Strategy    draft           45
                                  expired        122
                                  missed           3
                                  npo_rescheduled  7
                                  pbc_rescheduled  2
                                  pending         15
                                  published        7
            Messaging             applied          7
                                  cancelled        2
                                  completed       14
                                  draft            6
                                  expired          5
                                  pbc_rescheduled  1
                                  pending          2
                                  published        2
            Multimedia            applied          1
                                  cancelled        2
                                  completed        1
                                  draft            2
                                  expired          5
                                  pending          1
            Public Relations      applied          1
                                  cancelled        6
                                  completed        9
                                  draft            6
                                  expired         11
                                  npo_rescheduled  1
                                  pbc_rescheduled  2
                                  pending          3
                                  published        2
Length: 150, dtype: int64
```

```
d_group = d4.groupby(['group_slug', 'state']).size().reset_index()

d_group.to_csv("d_group.csv")

d_group2 = d4.groupby(['group_slug','name', 'state']).size().reset_index()

d_group2.to_csv("d_group2.csv")
```

Question 2

```
# Read data

data = pd.read_excel("C:\\ teradata\\session_export.xlsx")

data.head()
```

	id	created_at	updated_at	state	description	consultant_id	nonprofit_id	scheduled_for	time_slots	organization_id	conference_line_id	project_category_id	partner_organization_id	archived
0	102	2017-01-10 19:04:55.571	2017-01-10 19:05:13.720	cancelled	test.	NaN	2036	NaT	{"2017-01-13 16:00:00","2017-01-13 18:00:00",...	239	NaN	20.0	NaN	f
1	110	2017-03-29 14:36:50.901	2017-03-31 18:00:00.402	completed	We're creating new websites for a couple of ou...	130022.0	144136	2017-03-31 16:00:00	{"2017-03-29 20:00:00","2017-03-30 20:00:00",...	4218	5.0	20.0	NaN	f
2	60	2016-03-17 16:08:47.427	2016-03-30 14:52:12.539	cancelled	The TASH website is a Wordpress website that w...	11865.0	11682	2016-03-31 19:00:00	{"2016-03-28 15:00:00","2016-03-28 20:00:00",...	2153	1.0	20.0	NaN	f
3	95	2016-08-24 19:37:31.044	2016-09-06 15:00:01.829	completed	We want to increase our online sales of tea an...	129767.0	131196	2016-09-06 13:00:00	{"2016-08-30 14:00:00","2016-08-30 15:00:00",...	2933	2.0	10.0	NaN	f
4	53	2016-03-17 00:05:42.531	2016-03-24 16:00:01.854	completed	We are currently in the process of redesigning...	11412.0	9367	2016-03-24 14:00:00	{"2016-03-23 15:00:00","2016-03-24 14:00:00",...	1370	1.0	20.0	NaN	f

```
# Calculate the number of sessions
data['count'] = 0
data.head()
```

	id	created_at	updated_at	state	description	consultant_id	nonprofit_id	scheduled_for	time_slots	organization_id	conference_line_id	project_category_id	partner_organization_id	archived
0	102	2017-01-10 19:04:55.571	2017-01-10 19:05:13.720	cancelled	test.	NaN	2036	NaT	{"2017-01-13 16:00:00","2017-01-13 18:00:00",...	239	NaN	20.0	NaN	f
1	110	2017-03-29 14:36:50.901	2017-03-31 18:00:00.402	completed	We're creating new websites for a couple of ou...	130022.0	144136	2017-03-31 16:00:00	{"2017-03-29 20:00:00","2017-03-30 20:00:00",...	4218	5.0	20.0	NaN	f
2	60	2016-03-17 16:08:47.427	2016-03-30 14:52:12.539	cancelled	The TASH website is a Wordpress website that w...	11865.0	11682	2016-03-31 19:00:00	{"2016-03-28 15:00:00","2016-03-28 20:00:00",...	2153	1.0	20.0	NaN	f
3	95	2016-08-24 19:37:31.044	2016-09-06 15:00:01.829	completed	We want to increase our online sales of tea an...	129767.0	131196	2016-09-06 13:00:00	{"2016-08-30 14:00:00","2016-08-30 15:00:00",...	2933	2.0	10.0	NaN	f
4	53	2016-03-17 00:05:42.531	2016-03-24 16:00:01.854	completed	We are currently in the process of redesigning...	11412.0	9367	2016-03-24 14:00:00	{"2016-03-23 15:00:00","2016-03-24 14:00:00",...	1370	1.0	20.0	NaN	f

```
for i in range(len(data)):
    count = 0
    created_date = data.iloc[i,1]
    time_slot = re.split(',',data.iloc[i,8].replace("{","").replace("}","").replace('"',""))

for j in range(len(time_slot)):
        if time_slot[j]!='':
            session_date = datetime.strptime(time_slot[j],'%Y-%m-%d %H:%M:%S')
        if created_date > session_date:
            count=+1
        else:
            break

data.loc[i,'count'] = count
data.describe()
```

	id	consultant_id	nonprofit_id	organization_id	conference_line_id	project_category_id	partner_organization_id	count
count	1665.000000	623.000000	1665.000000	1665.000000	625.000000	1660.000000	0.0	1665.000000
mean	960.714715	119543.778491	154597.100901	6459.753754	3.075200	9.595181	NaN	0.003003
std	527.625466	64421.586744	42170.557673	2287.879252	3.108989	6.101411	NaN	0.054734
min	34.000000	6.000000	1304.000000	48.000000	1.000000	1.000000	NaN	0.000000
25%	515.000000	81861.500000	154469.000000	5428.000000	1.000000	4.000000	NaN	0.000000
50%	938.000000	148214.000000	168654.000000	7071.000000	2.000000	10.000000	NaN	0.000000
75%	1419.000000	167335.500000	176948.000000	8294.000000	4.000000	14.000000	NaN	0.000000
max	1931.000000	183262.000000	183524.000000	9329.000000	32.000000	21.000000	NaN	1.000000

```
len(time_slot)
```

```
1
```

```
from datetime import datetime
```

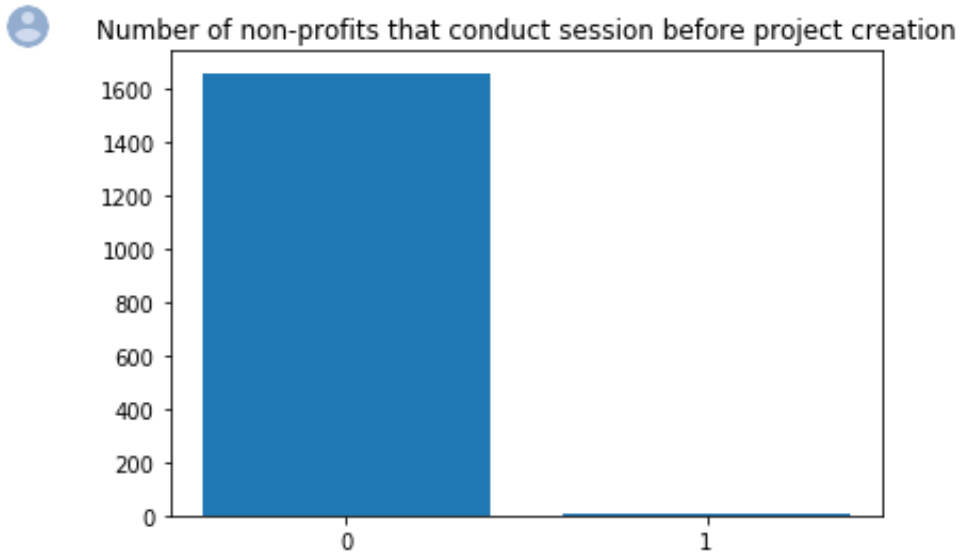
```
datetime.datetime(2017, 5, 12, 18, 30)
```

```
datetime.strptime(time_slot[1][1:20],'%Y-%m-%d %H:%M:%S')
```

```
# Number of nonprofits that conduct sessions before project starts
import matplotlib.pyplot as plt
DC = data.groupby(['count']).size().reset_index(name = "Pre_session_counts")
DC.head()
```

	count	Pre_session_counts
0	0	1660
1	1	5

```
plt.bar(DC['count'], DC['Pre_session_counts'], align = 'center')
plt.xticks(DC['count'])
plt.title('Number of nonprofits that conduct session before project creation')
plt.show()
```



Question 3

```
# Read data

import pandas as pd

from pandas import ExcelWriter

from pandas import ExcelFile

d1 = pd.read_excel("C:\\teradata\\project_export.xlsx",  sheet_name='project_export')

d2 = pd.read_excel("C:\\teradata\\project_categories.xlsx",  sheet_name='project_categories')

print("Column headings:")

print(d1.columns)

print(d2.columns)
```

```
Column headings:
Index(['id', 'organization_id', 'description', 'created_at', 'updated_at',
       'state', 'user_id', 'needs_accomplish', 'needs_support', 'needs_value',
       'campaign_id', 'image_id', 'project_inquiries_count', 'admin_id',
       'project_group', 'project_category_id', 'local_only', 'success_story',
       'partner_organization_id', 'match_job_id', 'satisfaction_rating',
       'agreed_at_community', 'timeline', 'publish_externally',
       'enable_success_story', 'is_archived', 'share_metadata'],
      dtype='object')
Index(['id', 'created_at', 'updated_at', 'group_slug', 'enabled',
       'international', 'name', 'slug'],
      dtype='object')
```

```
#Merge data

d1 = d1.iloc[:, [0,3,4,5,12,15]]

d2 = d2.iloc[:, [0,6]]

data = pd.merge(d1, d2, how = 'left', left_on = "project_category_id", right_on = "id")

data.head()
```

	id_x	created_at	updated_at	state	project_inquiries_count	project_category_id	id_y	name
0	5948	2017-07-06 17:27:44.906	2017-08-30 01:39:47.520	closed	3	19	19	Website design
1	5908	2017-06-28 16:55:47.853	2017-11-11 18:17:38.379	completed	3	14	14	Design
2	9339	2019-01-15 00:00:34.636	2019-01-15 00:00:34.636	draft	0	14	14	Design
3	5975	2017-07-10 15:36:19.184	2017-09-27 13:39:17.767	closed	0	2	2	Business Planning
4	9981	2019-04-03 03:19:57.849	2019-04-03 03:19:57.849	draft	0	1	1	Accounting & Finance

```
data = data.iloc[:, [0,1,2,3,4,7]]
```

```
data.rename(columns={'id_x': 'ProjectId'}, inplace = True)

# Time to complete each project by project type

data = data.query('state == "completed"')

data['Difference'] = data['updated_at'].sub(data['created_at'], axis=0).dt.days

data.head()
```

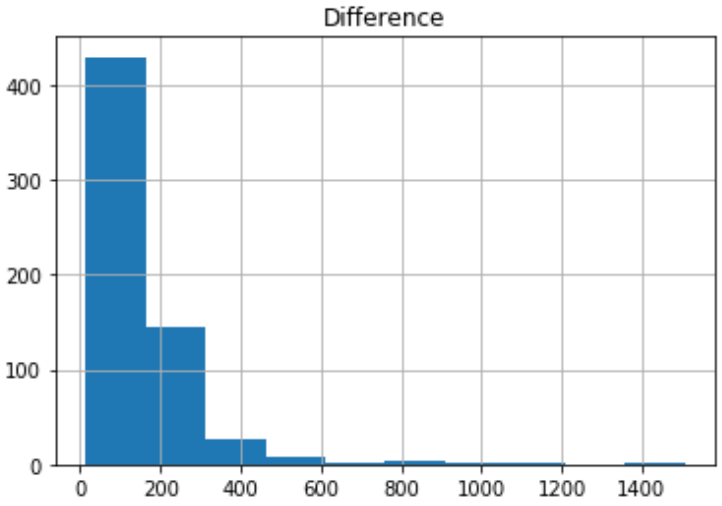
	ProjectId	created_at	updated_at	state	project_inquiries_count	name	Difference
1	5908	2017-06-28 16:55:47.853	2017-11-11 18:17:38.379	completed	3	Design	136
5	2646	2016-04-12 17:27:07.249	2016-10-31 17:44:05.194	completed	1	Marketing Strategy	202
10	5688	2017-06-03 13:01:34.481	2017-12-11 22:51:04.735	completed	1	Program Design	191
11	3588	2016-08-04 17:58:42.821	2017-08-28 15:04:48.039	completed	1	Business Planning	388
13	210	2014-11-25 00:11:37.809	2015-07-28 14:37:53.311	completed	1	Multimedia	245

```
import matplotlib.pyplot as plt

data.groupby("name")["Difference"].mean().sort_values()
```

```
name
Design                  151.701613
IT Infrastructure       156.800000
Accounting & Finance    158.331754
Staff Development       167.508772
HR Management           167.695946
Copy writing/editing    179.625000
Evaluation              183.629630
CRM                     188.137795
Brand Development       201.912791
Board Development       203.620000
Website development     205.009615
Messaging               207.539568
Research                208.346535
Website design          212.767528
Program Design          216.154762
Business Planning       219.415550
Multimedia              222.951613
Marketing Strategy      227.346789
Project Management      239.133333
Mobile Development      254.344828
Public Relations        256.896552
Name: Difference, dtype: float64
```

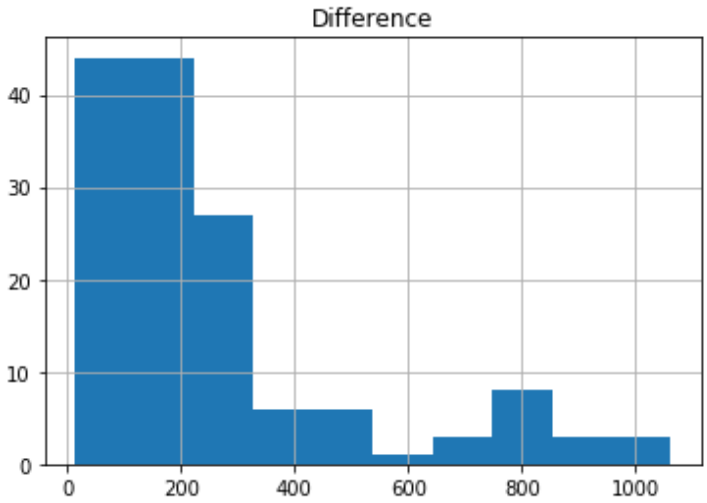
```
# Time to complete Design projects

data.query('name == "Design"').hist("Difference")
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000016703E40630>]],
      dtype=object)
```




```
# Time to complete Public Relations projects

data.query('name == "Public Relations"').hist("Difference")
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x00000167064C1EB8>]],
      dtype=object)
```




```
# Question 4

# Read data

import pandas as pd

from pandas import ExcelWriter

from pandas import ExcelFile

d1 = pd.read_excel("C:\\teradata\\project_inquiries.xlsx",  sheet_name='project_inquiries')
```

```
print("Column headings:")

print(d1.columns)
```

```
Column headings:
Index(['id', 'user_id', 'project_id', 'qualifications', 'created_at',
       'updated_at', 'state', 'time_slots', 'scheduled_for',
       'decision_deadline', 'conference_line_id', 'hours', 'pbc_rating',
       'npo_rating', 'satisfaction_rating', 'pbc_review', 'archived'],
      dtype='object')
```

```
d2 = pd.read_excel("C:\\teradata\\project_export.xlsx", sheet_name='project_export')

d2.columns
```

```
Index(['id', 'organization_id', 'description', 'created_at', 'updated_at',
       'state', 'user_id', 'needs_accomplish', 'needs_support', 'needs_value',
       'campaign_id', 'image_id', 'project_inquiries_count', 'admin_id',
       'project_group', 'project_category_id', 'local_only', 'success_story',
       'partner_organization_id', 'match_job_id', 'satisfaction_rating',
       'agreed_at_community', 'timeline', 'publish_externally',
       'enable_success_story', 'is_archived', 'share_metadata'],
      dtype='object')
```

```
d2 = d2.loc[:,['id','state','created_at', 'updated_at']]

# Select projects with multiple (>=2) volunteers

data = d1.query('state == "completed" or state == "accepted" or state == "confirmed"')

new = data.groupby("project_id")['id'].count().reset_index()

new=new[new.id>=2]

new.rename(columns={'id':'Vol_count'},inplace=True)

# Merge project data and project inquiries data

new_data = pd.merge(new,d2,how='left',left_on = 'project_id',right_on = 'id')

new_data.head()
```

	project_id	Vol_count	id	state	created_at	updated_at	Difference
0	840	2	840	completed	2015-06-13 18:02:13.378	2015-08-05 14:13:20.901	52
1	862	2	862	completed	2015-06-18 21:48:00.373	2015-11-10 14:09:12.727	144
2	876	2	876	completed	2015-06-22 16:20:29.597	2016-05-18 20:27:28.343	331
3	954	2	954	completed	2015-07-09 16:14:07.679	2016-11-29 15:50:48.091	508
4	1190	2	1190	completed	2015-08-27 17:50:43.032	2016-03-21 20:31:46.671	207

```
# Compute the time difference (time to complete) of each project
new_data['Difference'] = new_data['updated_at'].sub(new_data['created_at'], axis=0).dt.days
new_data['Difference'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c8f61d1c88>
```