Using Machine Learning To Identify and Map Scrum Adoption

Ridewaan Hanslo

02 July 2019

1. Introduction

I. Background

Scrum is an Agile software development framework. It is the most adopted and successful Agile framework used globally. However, a large proportion of software projects continue to fail. Statistics taken from VersionOne.com indicate in the region of 90 percent of organizations that use an Agile methodology or framework implement Scrum as part of their solution.

Therefore could it be that the organizations that attempts to implement Scrum does not do so based on best practices? If they attempt to adopt Scrum are there indications that they will not successfully adopt Scrum? Is it possible to identify different states or cities that displays successful and unsuccessful Scrum adoption?

Our problem will identify South Africa's (SA) Agile organizations as the population sample. Our investigation will gather and analyze data from these organizations within the provinces in SA to explore the mapping of the Scrum adoption within the country.

II. Problem

Our solution is to use the data gathered from an online survey on organizations that currently use Scrum for their daily projects. The data will provide continuous variables that will identify the organizations rating on factors such as organization culture, management, teamwork, employee experience etc. The organizations also rated the adoption of Scrum within the organization i.e. have they adopted Scrum or not.

The data is cleaned, validated and verified using Exploratory Factor Analysis (EFA). Thereafter, the data is used in a Multiple Linear Regression (MLR) statistical analysis technique as the machine learning model to be built. The data model will be trained with 80 percent of the dataset and the remaining 20 percent will be the test data. Once the model has been tested and analyzed, we can identify which states or cities within South Africa indicates successful Scrum adoption and those that does not.

With the model built we use Foursquare and folium mapping to identify and inform organizations if a province (state) displays a higher adoption rate than others.

2. Methodology

I. Introduction

The following section discusses the methodology used to conduct the investigation. This section starts by discussing the Data Sources, followed by Data Cleaning, and Data Analysis.

II. Data Sources

As mentioned in the Section 1: Background, the population sample is taken from the South African organizations using Scrum. The data was gathered using an online survey questionnaire. The online tool used was Google forms and the data was made available via a csv download.

III. Data Cleaning

The data cleaning was very minimal for this project as the survey questionnaire was designed so that invalid responses or no responses would not allow the user to progress. The majority of the data cleaning that occurred was in the Jupyter notebook. The data cleaning was to prepare the data to be analyzed for the descriptive, exploratory, and regression analysis.

IV. Data Analysis

The analysis was broken down into 3 major sections. The first section looked at the descriptive analysis providing standard deviation, frequencies, and means of the Scrum adoption dataset. The second section looked at the exploratory factor analysis, and building a basic multiple linear regression (MLR) model that could be further expanded and built upon. The third section analyses the adoption dispersion across the country and plots the adoption results across the countries provinces.

3. Results

I. Introduction

The following section provides the results of the analysis conducted for the project. The subsections are broken up into the demographics, descriptive statistics, EFA, MLR, and Scrum adoption mapping.

II. Demographics

The demographics of the survey respondents have been broken up into Province, Job Title, and Age Group.

Figure 1 displays the respondents based on the province (state) from which they reside. The majority of the respondents reside in the Gauteng province with 49.8% of the population sample.

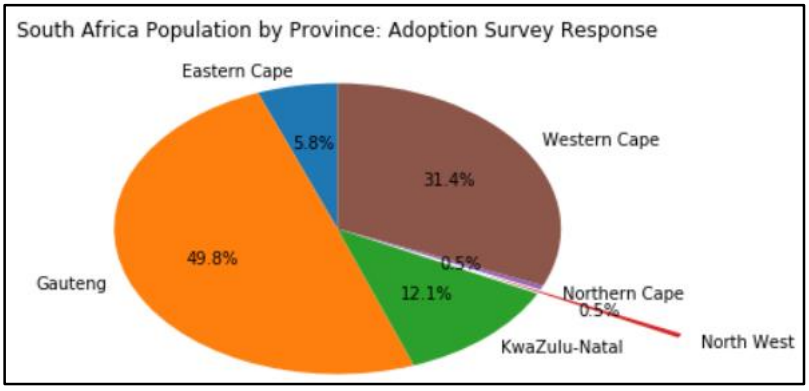

***Figure 1**: South African Population by Province*

Figure 2 depicts the job title of the respondents. The author was pleased to see that a variety of job titles have been recorded. It allows the project to have a greater variety of perspectives coming from organizations. The Scrum Master title recorded 19.8% of the responses.

The age group of the population sample ranged from 18-20 years to 39-59 years. The 29-38 years age group category and the 24-28 years age group had a combined percentage of 76.8%. Figure 3 displays the age group frequencies.

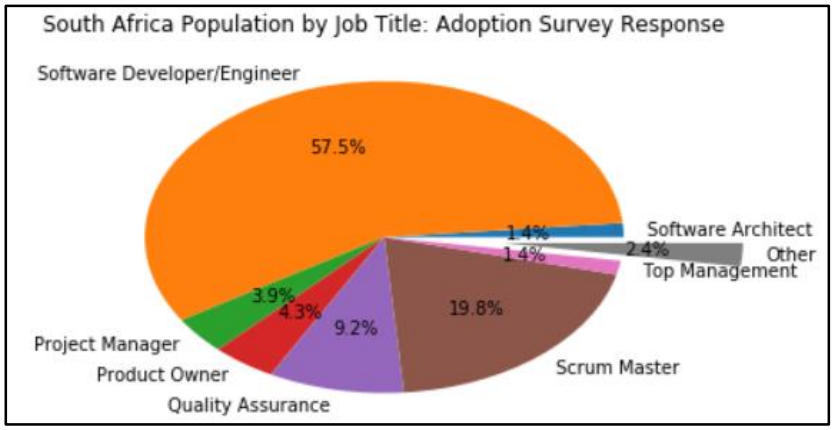

***Figure 2**: South African Population by Job Title*

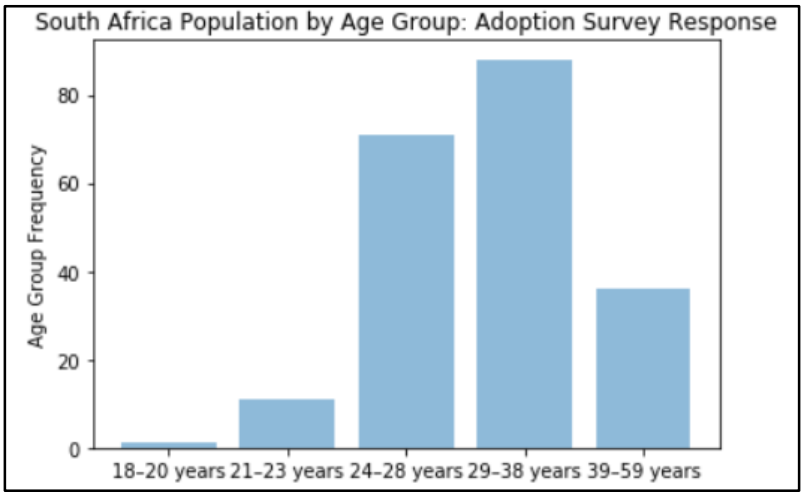

***Figure 3:** South African Population by Age Group*

III. Descriptive Statistics

Figure 4 displays the basic descriptive statistics that are derived from the responses.

	WorkWithinSA_A1	ScrumUsage_A2	AgeGroup_A3	Province_B1	JobTitle_B2	WorkExperience_B3	WorkWithinOrg_B4
count	207.0	207.000000	207.000000	207.000000	207.000000	207.000000	207.000000
mean	1.0	4.420290	4.710145	4.932367	5.594203	4.743961	3.125604
std	0.0	1.132989	0.832031	2.856651	14.832383	1.313540	1.485748
min	1.0	2.000000	2.000000	1.000000	1.000000	1.000000	1.000000
25%	1.0	4.000000	4.000000	3.000000	2.000000	4.000000	2.000000
50%	1.0	5.000000	5.000000	3.000000	2.000000	5.000000	3.000000
75%	1.0	5.000000	5.000000	9.000000	5.000000	6.000000	4.000000
max	1.0	7.000000	6.000000	9.000000	99.000000	7.000000	7.000000

***Figure 4:** Descriptive Statistics of the project*

IV. Exploratory Factor Analysis

Basic EFA analysis has been done on the data to find the factors of significance on the limited survey data recorded. Of the responses a basic SPSS EFA analysis has been generated to display the factors derived from the survey variables. Table 1 displays the KMO and Bartlett's test results. Figure 5 displays the Scree Plot of the EFA results. In a more elaborate research opportunity, the author can dig deeper on the EFA analysis.

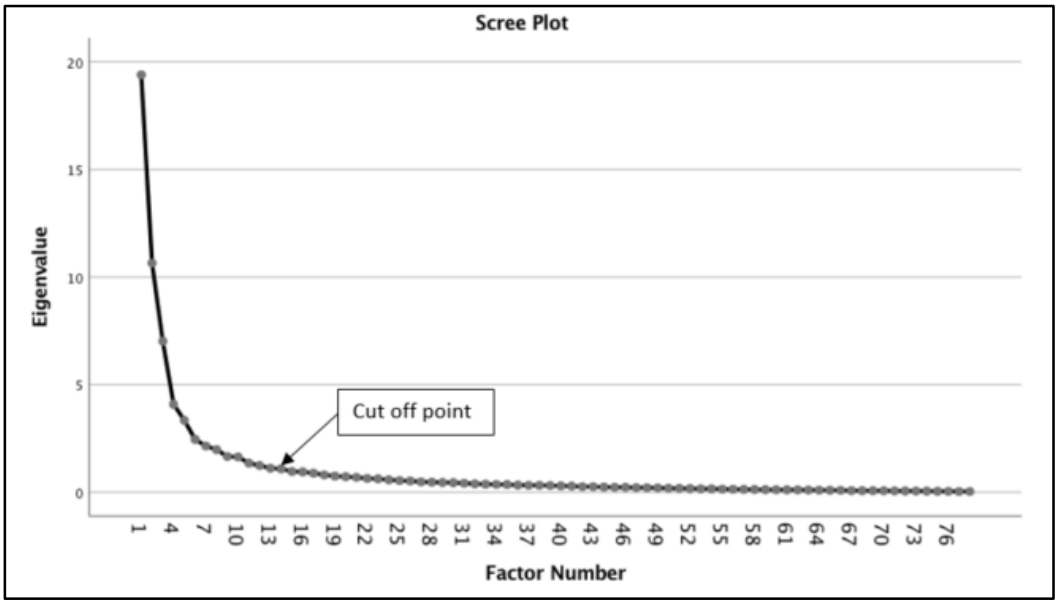

Figure 5: *Scree Plot for the Factors.*

Table 1: *KMO and Bartlett's Test Results for the EFA.*

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.879
Bartlett's Test of Sphericity	Approx. Chi-Square	15765.511
	Df	3003
	Sig.	.000

V. Multiple Linear Regression Model

For the project, a machine learning model has been developed to train the data and test the model for its accuracy. The accuracy that has been achieved needs further development with a variance score between ~0.40 to ~0.80 accuracy. The author would advise further development of the model for future research. Figure 6 displays partial results from the coefficients generated during the development of the MLR model.

```
Coefficients:  [[ 0.2629548  -0.1718043  -0.04751236 -0.19088155 -0.02120981 -0.01685441
   0.03589423 -0.00990481  0.07244454  0.08612357 -0.2187448   0.12017564
   0.06509107 -0.10111133  0.21034699  0.03975154 -0.0890368  -0.12790865
   0.10027975  0.01105817 -0.02143907  0.15537124 -0.11637622 -0.11546516
  -0.05987753  0.10319157 -0.11755821  0.06521124 -0.04029331  0.16922373
  -0.18814219 -0.06201037 -0.19195809  0.17046082  0.3264677  -0.33218474
  -0.05486219 -0.05714516  0.05818345  0.19909266 -0.12107272  0.06207647
  -0.15043556 -0.03228923 -0.05517595  0.08404214  0.17129823 -0.21190744
   0.01057313 -0.00549803 -0.10445267  0.15882121  0.06909275  0.20975819
  -0.01171534 -0.32892371  0.30253181  0.05731448 -0.10796836 -0.1716132
   0.12520852  0.16267298 -0.35043984  0.00587423 -0.13991979 -0.09221308
   0.26572039 -0.00343887  0.32798381  0.03161179 -0.22868844  0.0543215
   0.2932309   0.39826162 -0.0269742  -0.09684351  0.10509685  0.10696488]
```

***Figure 6**: Coefficients of the MLR machine learning model.*

VI. Scrum Adoption Mapping

The mapping of the adoption within the South African organization community context will be discussed next. The major output of this research project is to map the adoption results to the South African landscape broken up into the neighborhoods.

To plot the data the address details of the responses had to be converted into geo-coordinates.

The Google API was used to convert the addresses into coordinates. Figure 7 displays the code and output of the conversion.

```
# Extract the coordinates
pd.set_option('mode.chained_assignment', None) # set the warning off

for i in range(0,len(df_adoption_renamed)):
    df_adoption_renamed['Latitude'][i],df_adoption_renamed['Longitude'][i]=geocode_address(df_adoption_renamed.iloc[i]['ConcatenatedList'])
```

```
# Create function for extrating coordinates from google api

def geocode_address(loc):
    geocode_result = gmaps.geocode(loc)
    latitude = geocode_result[0]["geometry"]["location"]["lat"]
    longitude = geocode_result[0]["geometry"]["location"]["lng"]
    return latitude,longitude
```

ProvinceMappingToB1	City	Suburb	ConcatenatedList	Latitude	Longitude
Eastern Cape	Port Elizabeth	Adcockvale	Adcockvale, Port Elizabeth, Eastern Cape	-33.9417	25.5814
Eastern Cape	Port Elizabeth	Algoa Bay	Algoa Bay, Port Elizabeth, Eastern Cape	-33.9608	25.6022
Eastern Cape	Port Elizabeth	Amsterdam Hoek	Amsterdam Hoek, Port Elizabeth, Eastern Cape	-33.8546	25.6172
Eastern Cape	Port Elizabeth	Arcadia	Arcadia, Port Elizabeth, Eastern Cape	-33.8646	25.5119
Eastern Cape	Port Elizabeth	Benkamma	Benkamma, Port Elizabeth, Eastern Cape	-33.9622	25.4919

***Figure 7:** Google Maps API used to convert addresses to coordinates.*

Once the coordinates have been generated the Foursquare API and folium packages have been used to map the Scrum adoption across the South African landscape. The adoption landscape has been broken up into 6 clusters, which represents the 6 provinces from which the responses have been derived. Figure 8 displays the mapping of the adoption cases.

Figure 8: *Mapping of the adoption cases.*

The above map displays the results with the machine learning clusters identified by the fill color of the circles. The reader will notice that there is 6 distinct fill colors, each representing a cluster. The border color is one of 2. The green color represents Scrum adoption and the red border color symbolizes Scrum rejection. Figure 10 gives a zoomed in diagram of the Western Cape Province with the spread of adoption and rejection statistics.

The results display the neighboring suburbs and their adoption statistics. Figure 9 displays the adoption and rejection statistics of the project.

ProvinceMappingToB1	Adoption	Rejection
Eastern Cape	10	2
Gauteng	88	15
KwaZulu-Natal	17	8
North West	0	1
Northern Cape	0	1
Western Cape	53	12

Figure 9: *Adoption Statistics*

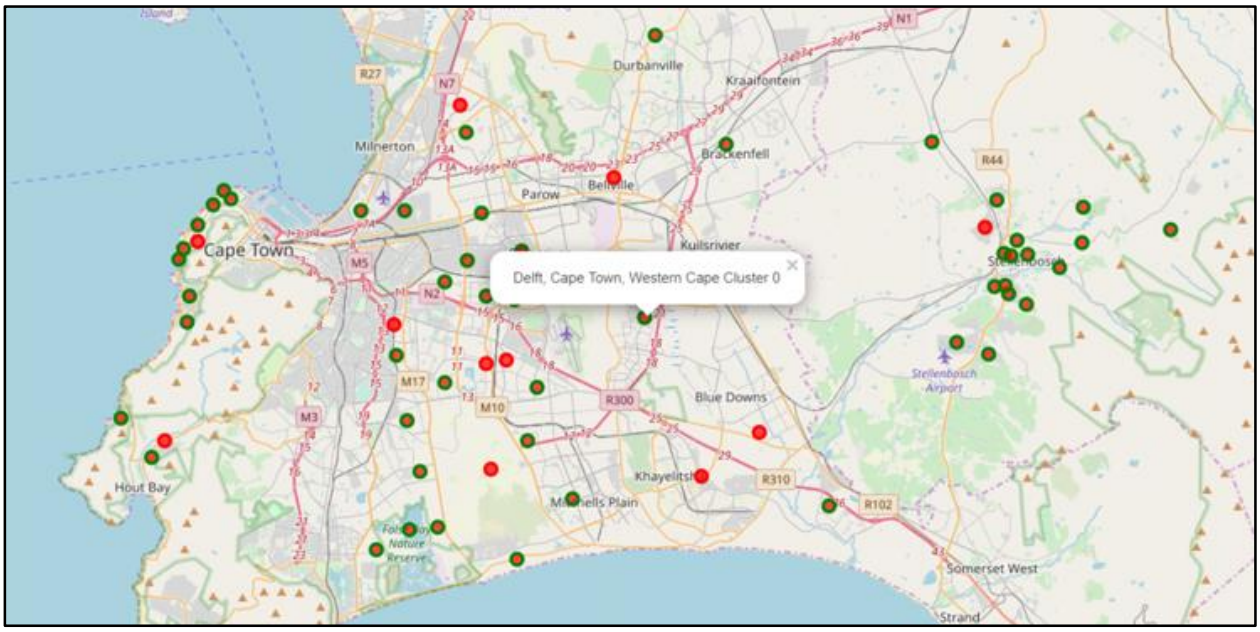

Figure 10: *Adoption results within the Western Cape Province.*

4. Discussions

The project has allowed the author to perform an online survey, followed by the analysis and display of the results. The findings from the results indicates that Scrum adoption rates is successful with more than 81 percent of organizations adopting Scrum within South Africa (SA). The province with the highest rejection of Scrum is the KwaZulu-Natal province with a 47% rejection rate, followed by the Western Cape Province with 23% rejection statistics.

The machine learning model that has been built allows individuals to test their adoption using the MLR statistical analysis technique. The author has made it known that the model required further training and testing for validation and verification accuracy. The model generates variable accuracy between ~0.40% and ~0.80%.

The demographics of the population sample signifies the interest of the greater SA community in the Agile adoption. Of the 9 provinces in SA, 6 of them captured responses.

5. Conclusions

Scrum being the most used and under researched Agile framework needs more research to be conducted. The research that is lacking is quantitative research. This project therefore looked at doing a quantitative study to identify the adoption rates and mapping it onto the SA landscape. The limitations of the study is the need for a larger population sample and the need to conduct additional Exploratory Factor Analysis and logistic regression to refine the analysis.

The author would also like to further refine the model with the assistance of a much larger dataset. Once the model has been refined with a much higher predictive rate, the author would like to proceed with a global survey and analysis study.