Context

This document contains analysis done for a Junior Business Analyst role for a company I applied to in Malaysia. The aim of the assessment was to select employees needed for an upcoming project based on a set of given employee data. The datasheet showed multiple employees with certain information including number of projects participated, type of project and accuracy of the type of project worked on. The aim of the assessment was to identify ten annotators to onboard by providing an explanation, thought process, and methodology of:

1. How the annotators where selected
2. Why these annotators?

To tackle the assessment, I used Microsoft Excel to clean and filter the data based on my preference and Tableau for visualization purposes.

Please note: The company (Appletoneuse), the CEO (Mark), and the employees are all **fictional.** The Tableau visuals and the data from Microsoft Excel are not real-time. They are just an aid to understand the project requirements better.

Table of Contents

Part 2: Annotator Onboarding 3

Selection process 4

BBox /Polygon projects joined by their AOW by their accuracy 7

80% - 90% bin 8

90% - 100% bin 13

Conclusion 15

Part 2: Annotator Onboarding

Mark, the CEO at Appletoneuse, has assigned SUPA to a project with the objective of detecting and counting apple growth in orchards. As with any client, we are going to provide Appletoneuse our best crop of annotators.

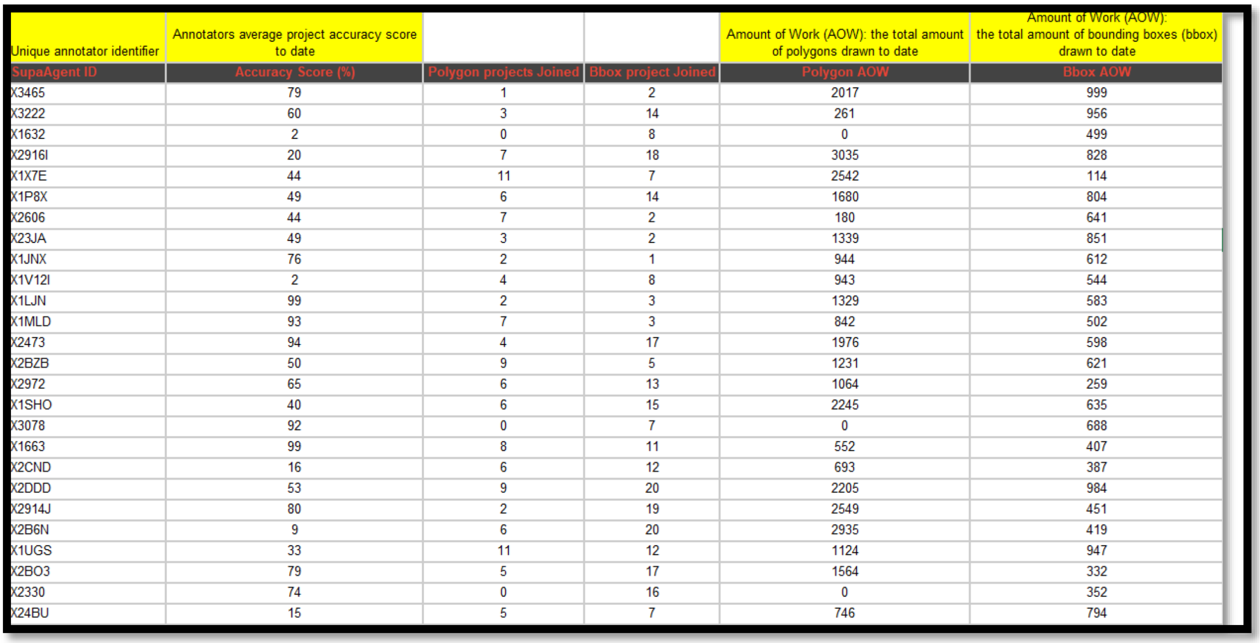
A rigorous selection process will be conducted with the main aim being choosing the right annotators guaranteed to provide results of high quality and in a timely manner. The annotators are chosen from an annotator pool based on several variables, namely:

- SupaAgent ID (annotator)
- Nationality
- Education
- Occupation
- Birth Date
- Date joined
- Accuracy Score (%)
- Polygon projects joined
- Bbox projects joined
- Polygon AOW (total amount of polygons drawn to date)
- Bbox AOW (total amount of bounding boxes (bbox) drawn to date)

A number of these variables will be disregarded because they have little effect on selection process. The only variables that'll play 'decisive' roles are **Accuracy Score (%)**, **Polygon projects** joined, **Bbox projects** joined, **Bbox AOW** and **Polygon AOW** and of course the Annotator (SupaAgent ID) for identification purposes.

A pool of annotators has been provided to us by HR. It will undergo strict conditional formatting based on a set of rules. These rules will help shorten the list, allowing us to choose the best from the pool.

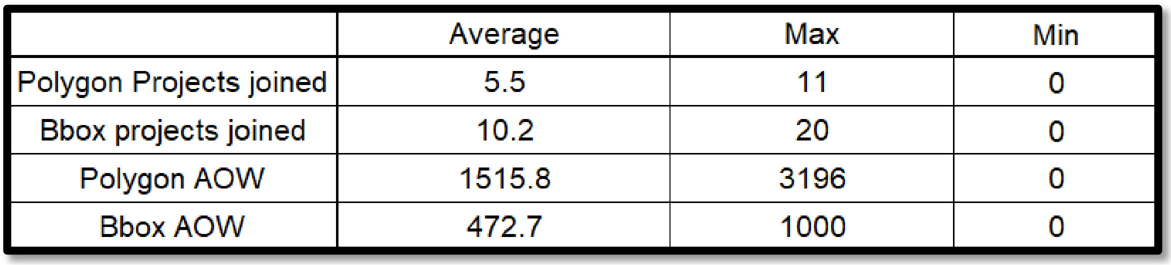
Selection process



Unique annotator identifier	Annotators average project accuracy score to date			Amount of Work (AOW): the total amount of polygons drawn to date	Amount of Work (AOW): the total amount of bounding boxes (bbox) drawn to date
SupaAgent ID	Accuracy Score (%)	Polygon projects Joined	Bbox project Joined	Polygon AOW	Bbox AOW
X3465	79	1	2	2017	999
X3222	60	3	14	261	956
X1632	2	0	8	0	499
X2916I	20	7	18	3035	828
X1X7E	44	11	7	2542	114
X1P8X	49	6	14	1680	804
X2606	44	7	2	180	641
X23JA	49	3	2	1339	851
X1JNX	76	2	1	944	612
X1V12I	2	4	8	943	544
X1LJN	99	2	3	1329	583
X1MLD	93	7	3	842	502
X2473	94	4	17	1976	598
X2BZB	50	9	5	1231	621
X2972	65	6	13	1064	259
X1SHO	40	6	15	2245	635
X3078	92	0	7	0	688
X1663	99	8	11	552	407
X2CND	16	6	12	693	387
X2DDD	53	9	20	2205	984
X2914J	80	2	19	2549	451
X2B6N	9	6	20	2935	419
X1UGS	33	11	12	1124	947
X2BO3	79	5	17	1564	332
X2330	74	0	16	0	352
X24BU	15	5	7	746	794

Figure 1. Employee datasheet

The image above is a sample image of the employee datasheet. The first action undertaken is to get a statistical summary of the above variables from the annotator pool:



	Average	Max	Min
Polygon Projects joined	5.5	11	0
Bbox projects joined	10.2	20	0
Polygon AOW	1515.8	3196	0
Bbox AOW	472.7	1000	0

Figure 1. Statistical annotator summary

At a quick glance, the determining factors at which the annotators are to be chosen is made a lot easier. These numbers will set the bar and no annotator below this bar does not qualify. **Mind you, the current stats include all the annotators, even those with the lowest accuracy scores**. The motive here is to get an overall summary of how the group is doing numerically.

There are 1,825 annotators to choose from with most not likely to make the final cut. To help simplify the election process, the annotators will be placed into two separate bins. The data sheet that contains the data will be subject to the following conditional formatting rules:

- Only annotators with accuracies between 80% - 100% will be considered
    - 4 annotators from **80% - 90%** will be chosen
    - 6 annotators from **90% - 100%** will be chosen

- Numbers pertaining to the above statistics will be the starting point:
    - Polygon Projects joined will commence at 5.5
    - Bbox Projects joined will commence at 10.2
    - Polygon AOW will commence at 1515.8
    - Bbox AOW will commence on 472.7

Settling for the above-mentioned accuracy scores has minimised the number of users from 1,825 to 394 from both bins. This makes the selection process easier and smoother.

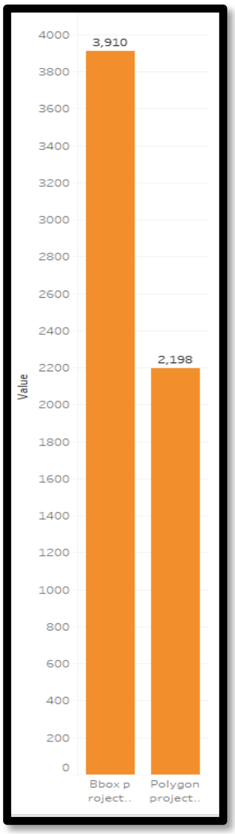


Figure 2. Bbox & Polygon project involvement

Another glance at summary in Figure 1 shows a notable difference between Polygon and Bbox annotations as well as Figure 2. Annotators seem to prefer Bbox projects (3,190) over Polygon projects (2,198). **This has set another task as to whether choose annotators who prefer Polygon or Bbox projects**. The figure alone does not provide concrete proof, Bbox projects could be popular but could have a lower accuracy score.

Data analysis will help uncover which project does better based on the accuracy score. The next step entails performing analysis comparing Polygon & Bbox projects against the Amount of Work (AOW) done in each project. Each of these variables will be analysed against their corresponding accuracy score.

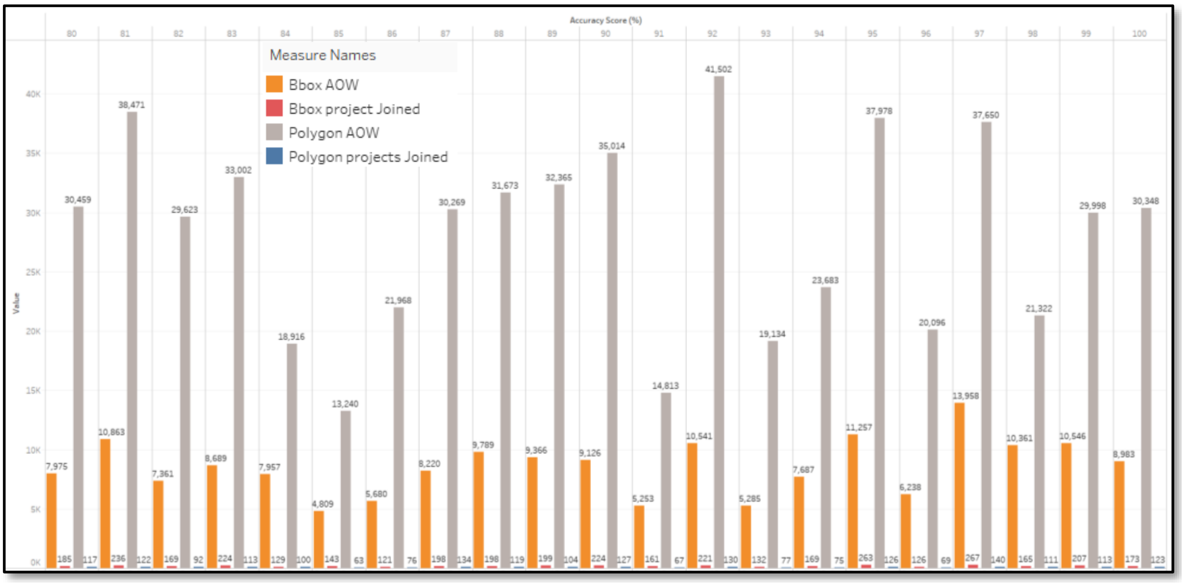


Figure 3. Cross comparison of annotators accuracy score by Polygon/Bbox AOW & Polygon/Bbox projects joined

BBox /Polygon projects joined by their AOW by their accuracy

As seen in Figure 3, Bbox projects (red bars) usurps Polygon projects (grey bars) in terms of project involvement, but when it comes to accuracy, the AOW is significantly higher for Polygons compared to Bbox. For example, at 89%, Polygon AOW is almost 4 times more than Bbox projects.

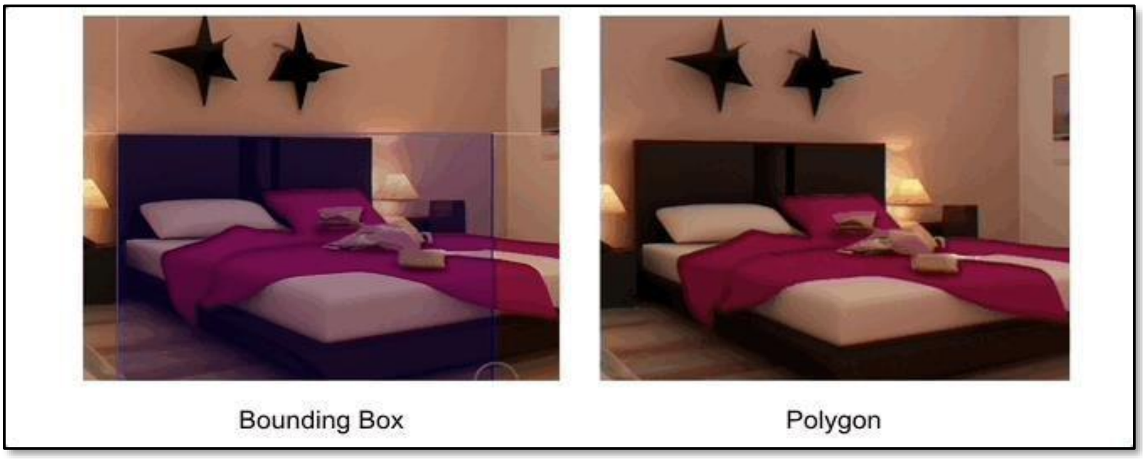
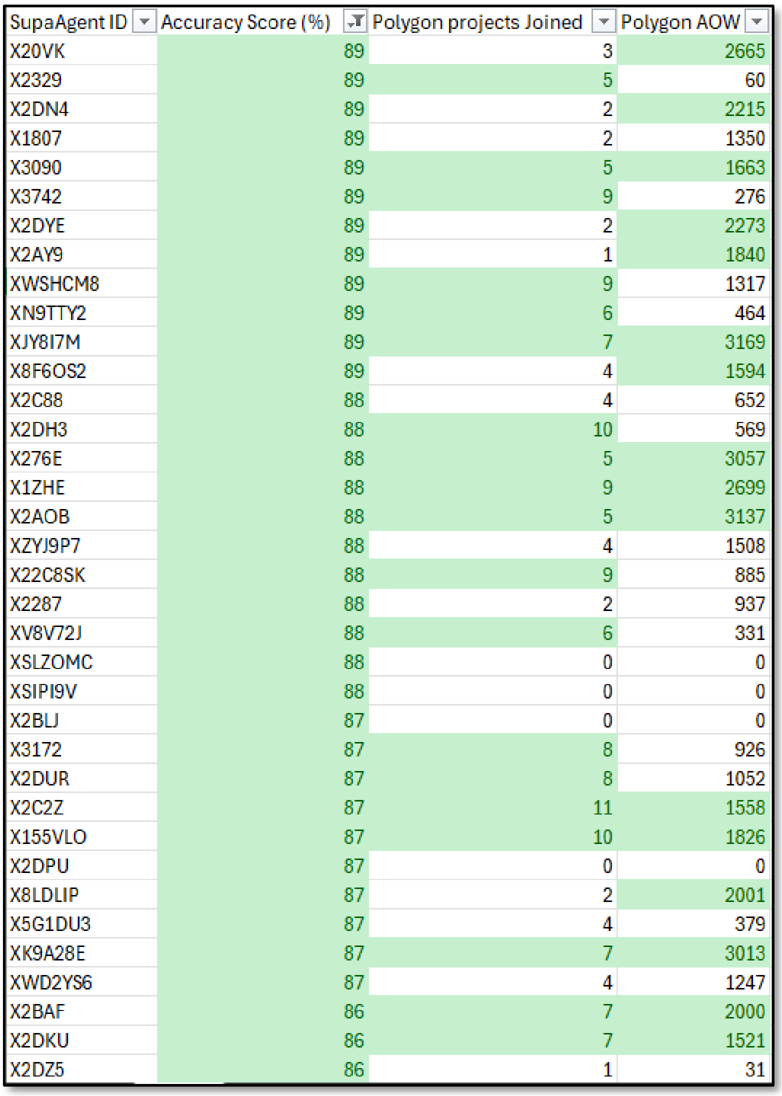


Figure 4. Bounding Box (Bbox) and Polygon illustration

Another simple explanation for why Polygon annotation will be chosen over Bbox for this project is down to bounding accuracy. As depicted from the above illustration, Bbox bounds the entire subject (bed) to include areas around the subject, whereas the polygon illustration only bounds the subject itself. The Bbox illustration has incorporated other aspects of the subject such as the floor, bed side lamp, wall etc., whereas the polygon only encapsulates the subject, which is the bed.

As a final decision, annotators will be chosen from the Polygon projects pool. Supporting this decision boils down to the AOW. There is a notable difference between each project at each accuracy score; however, the deciding factor was the AOW as the annotators are heavily involved in drawing objects and are exposed to different types of polygons.

80% - 90% bin



SupaAgent ID	Accuracy Score (%)	Polygon projects Joined	Polygon AOW
X20VK	89	3	2665
X2329	89	5	60
X2DN4	89	2	2215
X1807	89	2	1350
X3090	89	5	1663
X3742	89	9	276
X2DYE	89	2	2273
X2AY9	89	1	1840
XWSHCM8	89	9	1317
XN9TTY2	89	6	464
XJY8I7M	89	7	3169
X8F6OS2	89	4	1594
X2C88	88	4	652
X2DH3	88	10	569
X276E	88	5	3057
X1ZHE	88	9	2699
X2AOB	88	5	3137
XZYJ9P7	88	4	1508
X22C8SK	88	9	885
X2287	88	2	937
XV8V72J	88	6	331
XSLZOMC	88	0	0
XSIPI9V	88	0	0
X2BLJ	87	0	0
X3172	87	8	926
X2DUR	87	8	1052
X2C2Z	87	11	1558
X155VLO	87	10	1826
X2DPU	87	0	0
X8LDLIP	87	2	2001
X5G1DU3	87	4	379
XK9A28E	87	7	3013
XWD2YS6	87	4	1247
X2BAF	86	7	2000
X2DKU	86	7	1521
X2DZ5	86	1	31

With reference to the above image, after the data was subject to conditional formatting to filter out the annotators who have accuracy scores between 80% - 89%, the bin has reduced to 179 annotators from 392. For a much clearer explanation:

- Accuracies higher than **79%** are highlighted in green
- Polygons projects higher than **5.5** are highlighted in green
- Polygon AOW higher than **1515.8** are highlighted in green

It is important to note that those with a score of 89% are not guaranteed a place on the team. It is possible that an annotator with 88% has drawn more polygons and joined more projects or vice versa. While it's not difficult to do the analysis on an excel sheet, all these numbers and colours could cause confusion. Hence, Tableau will be used as the analysis tool to make tasks easier and much more visually appealing. The above excel sheet will be imported to Tableau for the required analysis and visualisations.

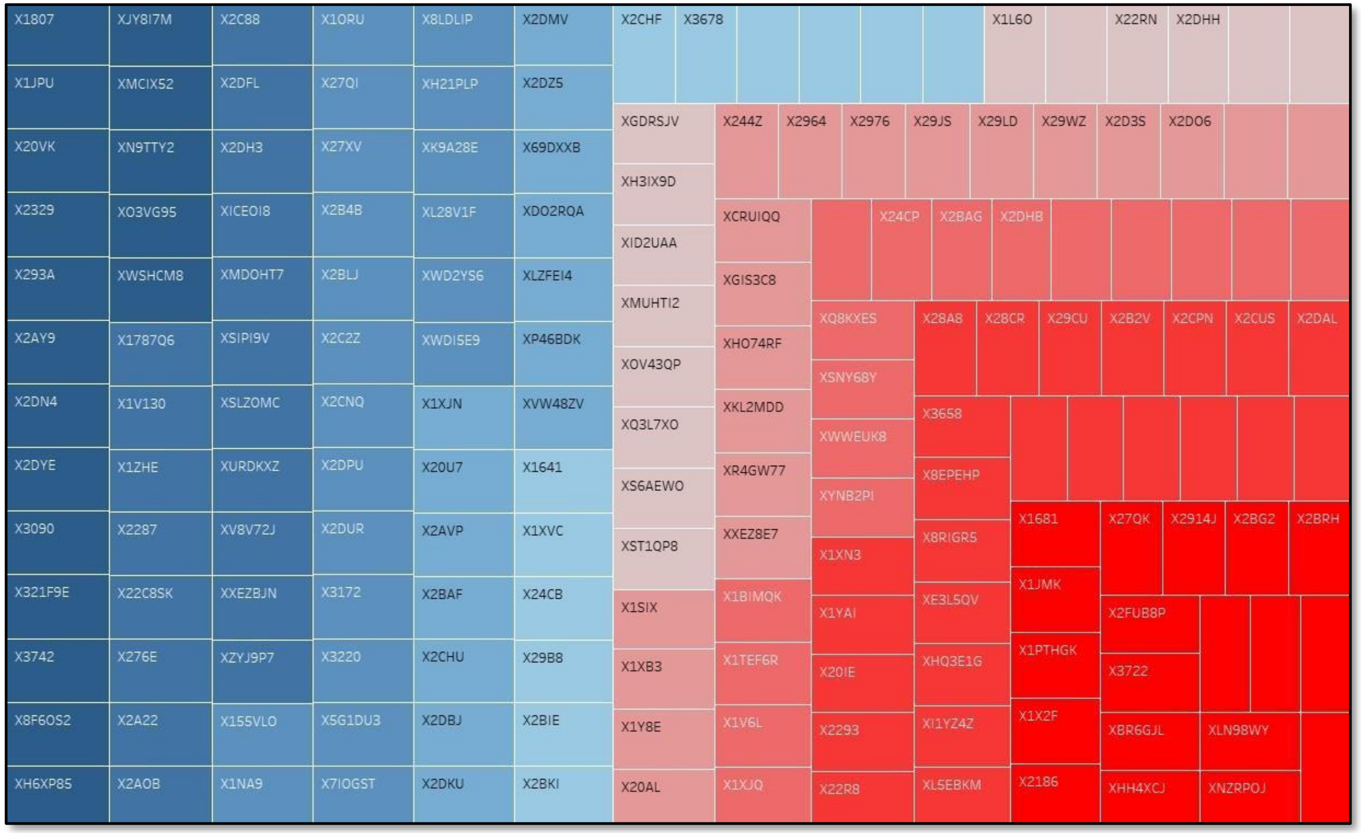


Figure 5. Accuracy Score (%) by Annotator

The above data displays a portion of the selected annotators with scores from 80% - 89% ordered largest to smallest, regardless of the project they've undertaken. The darker the shade of blue, the higher the accuracy; the darker the shade of red, the lower the accuracy. No decision can be made from this preview, but it provides a rough insight into the annotators chosen for this pool.

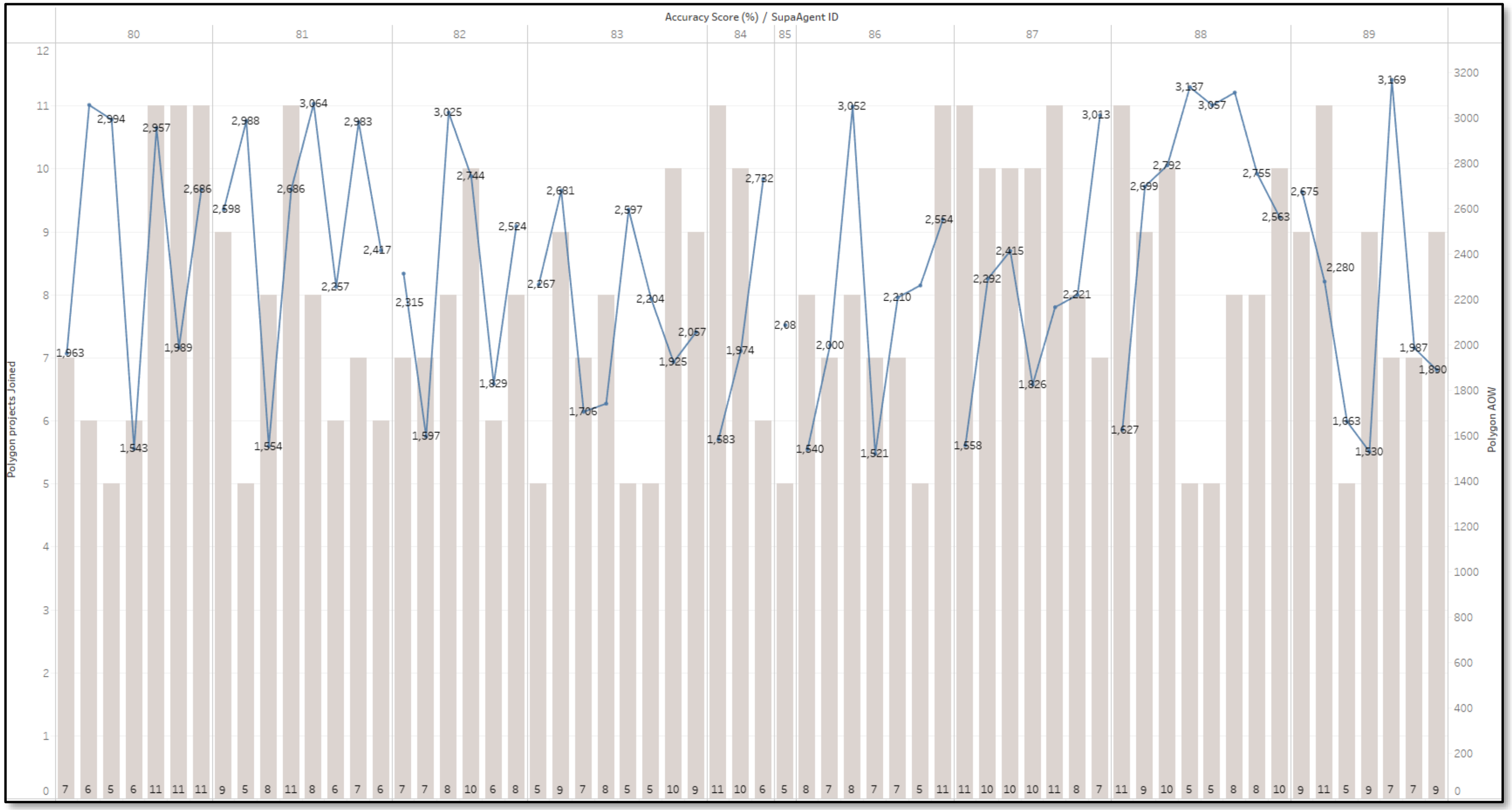


Figure 6. Comparison between Polygon projects joined, Polygon AOW & the Accuracy score of each annotator

After the data has been cleaned, filtered and imported, the selection process is made easier, as we can compare each annotator to each other. As previously mentioned, having a higher accuracy or more project involvements (grey bar) or more polygons drawn (blue line) does not **guarantee** a spot in the final list of annotators.

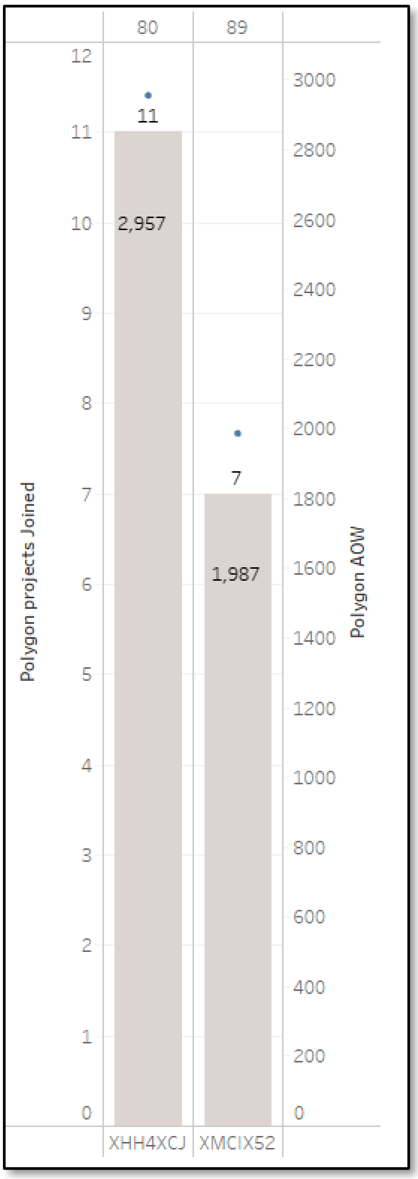


Figure 7. Annotator cross comparison

For example, the data above pertains to two annotators, XMCIX52 (89%) and XHH4XCJ (80%). The latter is involved in more projects (11), has a higher accuracy, and is exposed to more polygons (approximately a thousand more) compared to the former, who has a higher accuracy but lesser project involvement and AOW's.

Based on the data from Figure 7 and the logic above, the annotators will be chosen to best fit the project's purpose. For this bin only 4 annotators will be chosen. We must impress our client so that we can capture more projects in the future. The other 6 annotators will be chosen from the 90% - 100% bin. Since they have higher accuracy, they are more exposed to labelling data.

90% - 100% bin

This bin will also undergo the same process as the previous bin. To avoid redundancy, this section will begin with a cross-comparison analysis of annotators rather than re-explaining each process step-by-step. Before we proceed, this bin had a total of 392 annotators. However, after applying the specified filtering rules, the sample pool size has been reduced to 67.

Like the previous bin, maintaining a high accuracy does not guarantee a spot. An annotator could have impeccable accuracy but joined less projects or less AOW's. Data analysis will help simplify the selection process by weighing in the relevant factors.

(Next page)

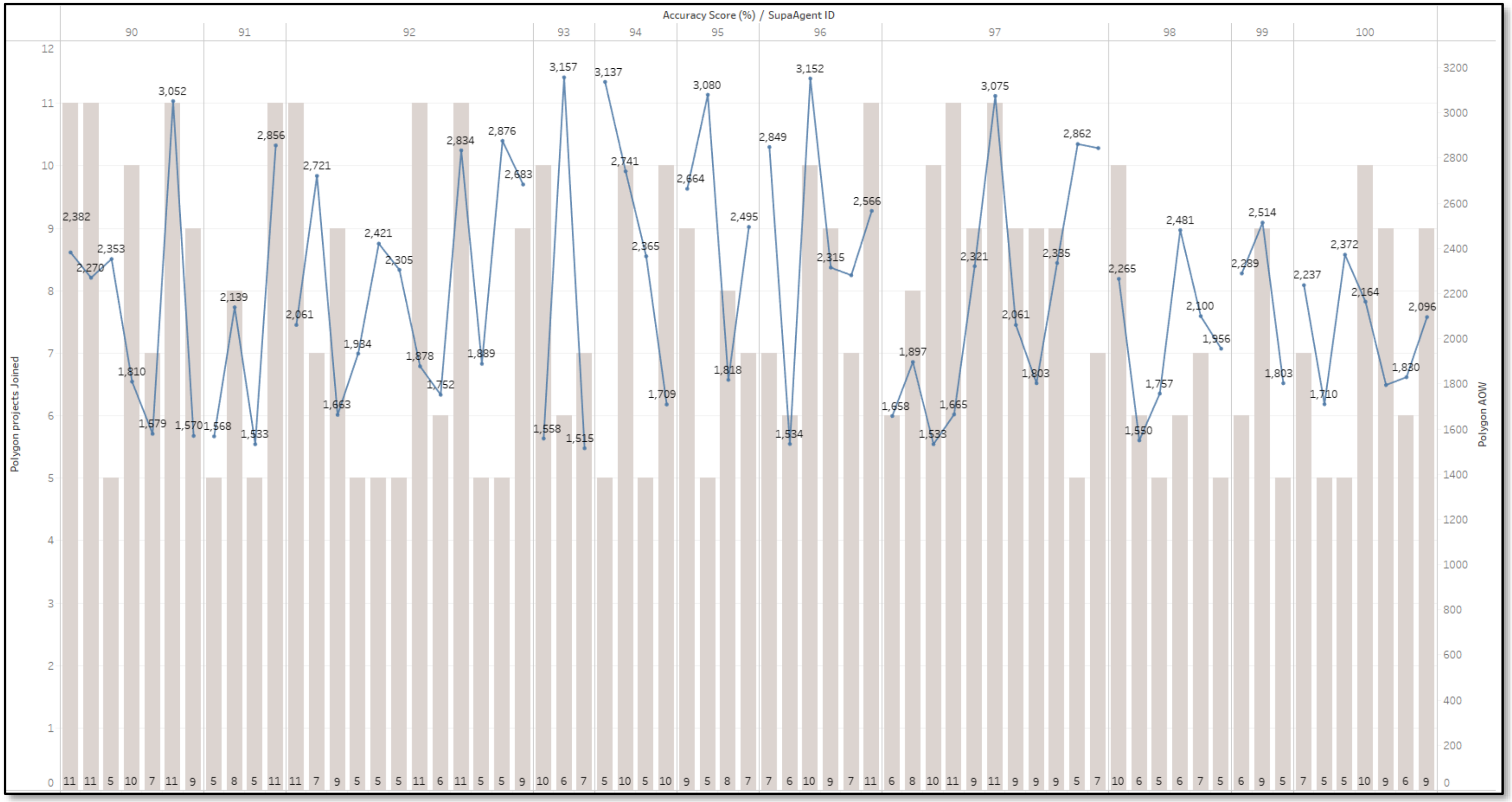


Figure 8. Comparison between Polygon projects joined, Polygon AOW & the Accuracy score of each annotator for 90% - 100% bin

Similarly, high accuracy stats do not confirm that the annotator will be chosen. From the above chart, the annotators that best fit the plan will be chosen to work for Mark.

Conclusion

This report documented thoroughly on the selection process of the annotators for the upcoming potential client, Appletoneuse. The process started by identifying the variables that adversely affect the selection process. Once those variables were identified, a statistical summary was obtained that would help when cleaning and filtering the data. The summary was used to create a set of rules in which the annotators would be chosen from.

After determining the type of annotator, the annotators were separated into two bins that would form the shell of the selection. The data sheet was then subject to conditional formatting based on the statistical data and the rules to further cleanse the pool to further simplify the selection process for each bin. Once all these steps were completed, the data was analysed using a visualization tool, Tableau, which facilitated the selection process. A series of visuals were employed to aid in the selection process, allowing a visual cross-comparison between each annotator to see notable differences. The final list of annotators to be sent to Mark is as follows:

80% - 89% bin			
SupaAgent ID	**Accuracy Score (%)**	**Polygon Projects**	**Polygon AOW**
XYJ8I7M	89	7	3,169
X2A0B	88	5	3,113
X2DBJ	86	8	3,052
XHQ3E1G	81	8	3,064
90% - 100% bin			
SupaAgent ID	**Accuracy Score (%)**	**Polygon Projects**	**Polygon AOW**
X23QW27	99	9	2,514
X233X	97	11	3,075
X15K7L8	96	10	3,152
X23AQ	95	5	3.080
XFIXDMF	93	6	3,157
XD31B4V	90	11	3,052

Table 1. Selected annotators

We could have chosen an equal number of annotators from each bin, but the idea here is to use those with more exposure and experience to teach their successors. This way, for the next project or any other prospective project, the annotators from the 80% - 89% bin will be the ones teaching their successors, thus allowing the cycle to keep repeating itself. Some annotators with lower accuracy scores but more project involvements were selected over those who have higher accuracy scores. The rationale behind their selection is that by the time they achieve a higher accuracy score, they would have already had more project involvement and exposure.