DSL21 Project: CERBERUS

Umut Ekin Gezer(03716498), ge36mil@mytum.de,

1-)Significant Influence Factors of the Simulation Time (Wall Clock Time):

The given csv file “simulation_time_data.csv” includes 33 columns, which are collected from experiment of Cerberus: The Power of Choices- Simulation. In this part, I will detect significant features in data frame, which have important influences on simulation time (wall clock time).

First the I calculate wall clock time every simulation by using “end time” and “start time” columns. The differences between of these columns give me simulation time. I merge the calculated “wall_clocktime” with simulation time data frame, then I delete start time and end time which are not anymore necessary factors on my research. My aim is detecting which features of data frame have significant influences on wall clock time. Secondly, I check every column property of data frame. Regarding of the data frame information we can already see that some features will not be relevant in the exploratory analysis since there are some missing values (such as nodes and fk_decorated_topology_id). In addition, there is so many features to analyze that it might be wise to concentrate on the ones which can give real insights. I remove Id and the features with 30% or less Nan values. In Figure 1, see how the wall clock time data is distributed. It is remarkable that the wall_clocktime data is right skewed and some outliers between 600.000s and 800.000s of data. However, based on data size (781) the outliers are not quite significant.

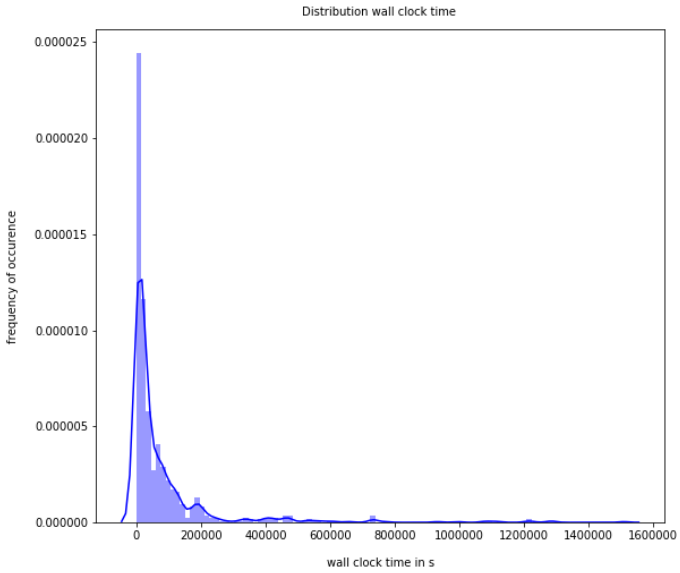

Figure 1: Distribution wall clock time

To detect significant factors of wall clock time, I will conduct correlation test (Correlation test can merely conducted by float or integer values), thus I extract the columns which have categorical object values. Furthermore, id and name columns refer the same type of data. For instance, “algorithm_id” and “algorithm_name” refer same parameters. After object and missing value extraction, the data frame consists of 14 columns which merely consist of integer and float variables. Based on the Figure 2, there are only one type of cpgen_id, finished, fvgen_id features. They have no influences on our analysis. I could have extracted them, but the correlation test will detect them. Thus, extraction of those parameters is not compulsory for the significance influence analysis.

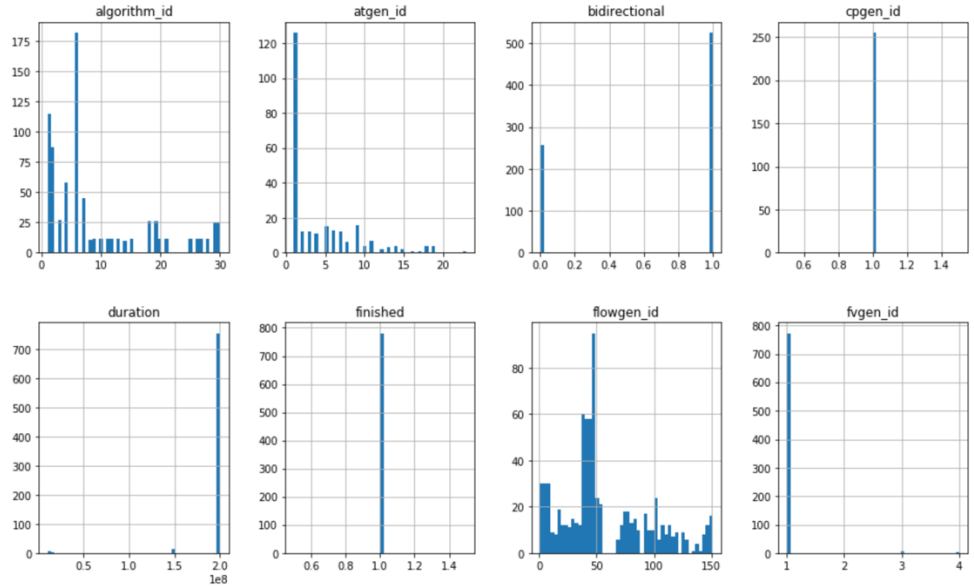

Figure 2 Numerical data

First, I conduct correlation analysis, which aims to detect the influences of the features on the target column of wall clock time. Among the features with numerical data, number of flows and arrival time generator id parameters have important correlation with wall clock time.

Their absolute correlation values with wall clock time data are greater than threshold value (0.5). I visualize numerical feature correlation with wall clock time in Figure 3.

Note: “Number of flow” is numerical data, increase on data has influence on wall clock time according to Figure 3 but the feature “atgen_id” seems to integer categorical variable. Nevertheless, the arrival time generator has strong correlation with on wall clock time.

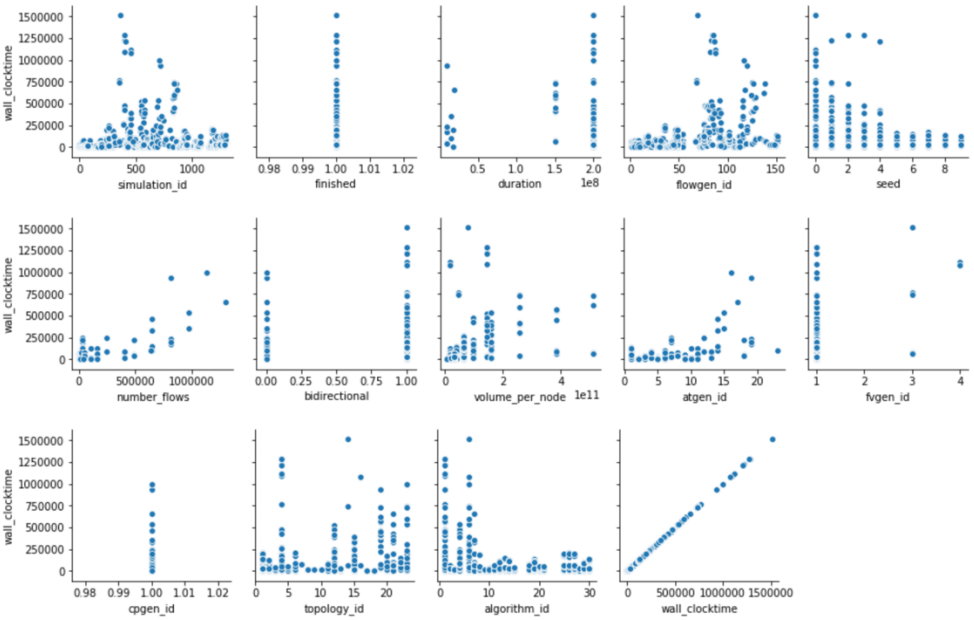

Figure 3 Numerical data correlation

However, correlation test might not always explain all relationship with feature and target. For instance, curvilinear relationship cannot be identified just by looking correlation values. Therefore, the proer test to detect significant influence factors is ANOVA (Analysis of Variance) analysis. Regard of ANOVA analysis, the p values of each feature are detected. If p values are smaller than threshold alpha value (in our case alpha=0.05) we can come the conclusion that belonging features are significant influence factors on target column “wall_clock_time”. I conduct the ANOVA test separated by the object features and numerical features to identify effectivity of previous correlation test for numerical features.

According to ANOVA test applied on the features of object values, “sha”, "simulation_behavior_params”, “flowgen_name”, “atgen_name”, ”atgen_parameter”, ”fvgen_parameter”, “topology_parameter”, ”algorithm_name”, ”algorithm_parameter” have significant influences on the target column “wall_clock_time”. For numerical features duration "flowgen_id" "number_flows" "bidirectional" "volume_per_node" "atgen_id" "fvgen_id" "topology_id" "algorithm_id" have significant influences on wall_clocktime.

As a conclusion for numerical feature in correlation test only number_flows and arrival time generator id (atgen_id) have correlation values greater than threshold value 0.5 . The correlation test does not only detect significant the columns which have significant influence of the target feaure "wall_clock_time", it detects additionally , from which columns, the linear regression can be done for target feature.

Tum sum up we can make predictions and build model for wall_clock time with the features arrival time generator (atgen_id) and number of flows (number_flows). But the features "flowgen_id", "number_flows" ,"bidirectional", "volume_per_node", "atgen_id", "fvgen_id" "topology_id" ,"algorithm_id", “sha”, "simulation_behavior_params”, “flowgen_name”, “atgen_name”,”atgen_parameter”,” fvgen_parameter”, “topology_parameter”, ”algorithm_name”, ”algorithm_parameter” have significant influences on wall_clocktime.

2-)Influence of Algorithm on Completion Time

One Way ANOVA test is a proper test to identify influence of more than 2 levels of interest and groups. In this part I will conduct ANOVA test in 'flowcompletion_data.csv' to identify impact of the feature algorithm_id on demand completion time which is the target feature column “completion_time”. In algorithm id there are 10 groups of algorithms.

The main idea of one way ANOVA test is take random samples from each group, then compare the sample means variation between the groups to the variation within the each groups. At the end decide based on statistical test for the means of groups equal or not. There are three conditions of ANOVA test. First one is samplinf random and independent variables. Second is cheching normality whether the population is normally distributed. I controlled second condition with Q-Q plot. The third one is homogenity of variance, which is satisfied if the ratio between largest and smallest sample standart deviation is not greater than threshold value(2).

Finally, if all three conditions are satisfied the one way ANOVA hypothesis test can be conducted. The two hypothesis are claimed. Hypothesis zero is all sample means are equal . Hypothesis not all sample means are equal. To detect which hypothesis is true the level of significance (alpha value) is specified. At the end the p-value approach or the critical value apprach (F-Test) might be conducted to detect which hypothesis is true. The zero hypothesis will be rejected if calculated p-value is smaller than alpha value or F value in Between Groups is greater than F critic value in Between Groups. The p and F values will be detected by statistical calculation and I will build an ANOVA table to test my hypothesis.

First I take 50 random sample from each simulation to detect influence of algorithm id on completion time. Thus first condition of ANOVA is satisfied. Then I divided the samples into the 10 groups of each algorithm id since my hypothesis is whether algorithm id has significant influences on completion time. In FIGURE 4, I visualize completion time and each algorithm id with box plots and in FIGURE 5 the distribution of completion time for each algorithm id.

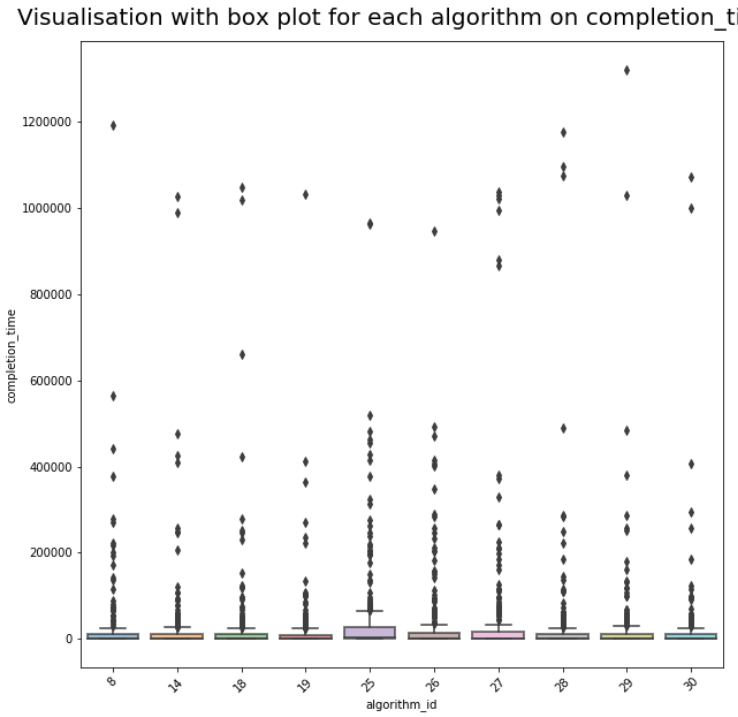

Figure 4 Algorithm on completion time.

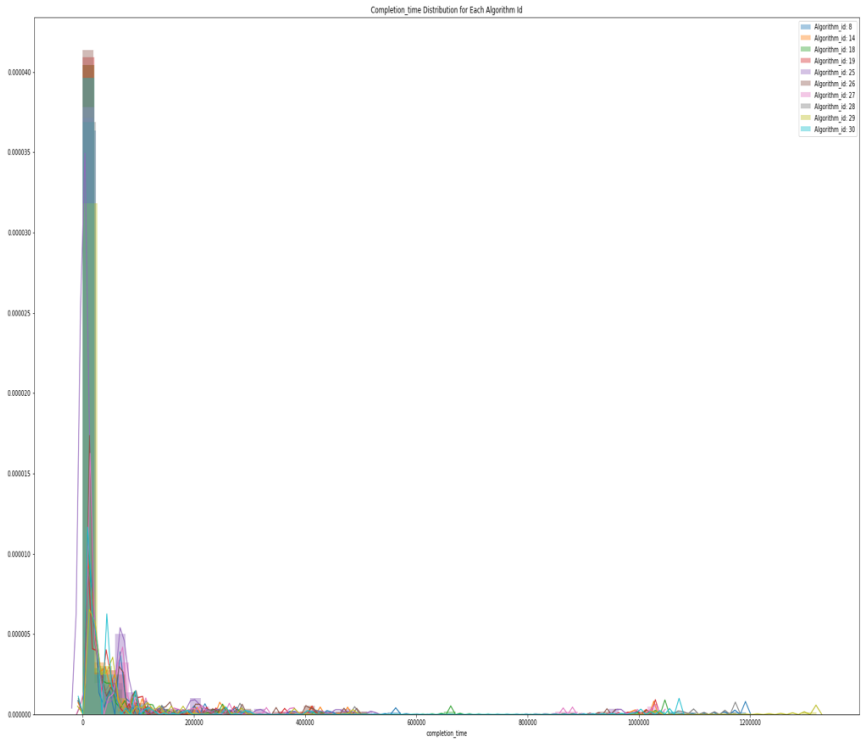

Figure 5: Completion time Distribution for Each Algorithm

To satisfy second condition of ANOVA, I plot Q-Q plots. The data for each group shows roughly strait line. There exist some of outliers in some plots, but I prefer to neglect them since the data is real data. The second condition of ANOVA is also satisfied.

To satisfy third condition I calculate ratio between maximum standard deviation and minimum standard deviation. If the ration smaller than 2, then the assumption of homogeneity of variance will be fulfilled. The ratio of the largest to the smallest sample standard deviation is 1.96, The threshold value is 2, Thus third assumption is fulfilled. Then All 3 conditions for ANOVA are fulfilled. Since all conditions of ANOVA are satisfied, the hypothesis can be claimed.

Zero hypothesis is “All sample mean completion time of each algorithm ids are equal”. Hypothesis one is “Not all completion time means are equal”. I define the level of significance (alpha value) 0.05. Then I can build the one way ANOVA table. Inside the ANOVA table differences between groups is called Treatment(TR) and difference within the groups is called

Error(E).First I identify degrees of freedom which is calculated based on count of groups.Then I calculated the SSTR(sum of squares due to treatment) and SSE(sum of squares due to error) and their sum SSTO based on statistical calculation. Figure 6 shows the ANOVA table of my research.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2.043e+11	9	2.27e+10	2.08911	0.0272661	2.11789
Within Groups	3.2489e+13	2990	1.08659e+10			
Total	3.26933e+13	2999	1.09014e+10			

Figure 6: ANOVA Table

Finally I conduct p-value approach and the critical value approach. Since the p-value PR(>F) i s less than our error rate (0.05 > 0.010566), we could reject the null hypothesis. This means w e are quite confident that there is a difference in completion_time sample mean for each algori thm_id. Secondly the according to critical value approach, the F-score is 2.089114098795684 and the critical value is 2.1178938207595115. This also means we are quite confident that the re is a there is a difference in completion_time sample mean for each algorithm_id. For both te st, we failed to reject zero hypothesis.

As a conclusion it is strong evident that not all average completion time are the same for different algorithm id, at 5% significance level. Thus, the algorithm_id has no significant impact on completion_time.

3-)Confidence Interval of Completion time each Simulation Id and Algorithm ID groups:

In this part I calculated 95% confidence interval for completion time and relevant values such as sample and population mean, sample and population standard deviations for each multiple groups of algorithm id and simulation id which are taken from csv file 'flowcompletion_data.csv'. The confidence interval specifies range of values, which gives the probability of expected true population parameter lies in.

Before I start to calculate relevant population and sample values, I take 100 random sample from each algorithm and simulation id groups. There are 30 groups which are combination of algorithm and simulation id groups. Then I calculated for all population in data frame mean, median, standard deviation of completion time. I conduct the same operation of each group to calculate their confidence interval for completion time. I collect every finding sample in a data frame "result". I visualize the distribution of sample means for completion time in Figure 7. The distribution of sample means looks left skewed. But no outliers. The simple samples are collected randomly from each of groups, and success failure condition is satisfied.

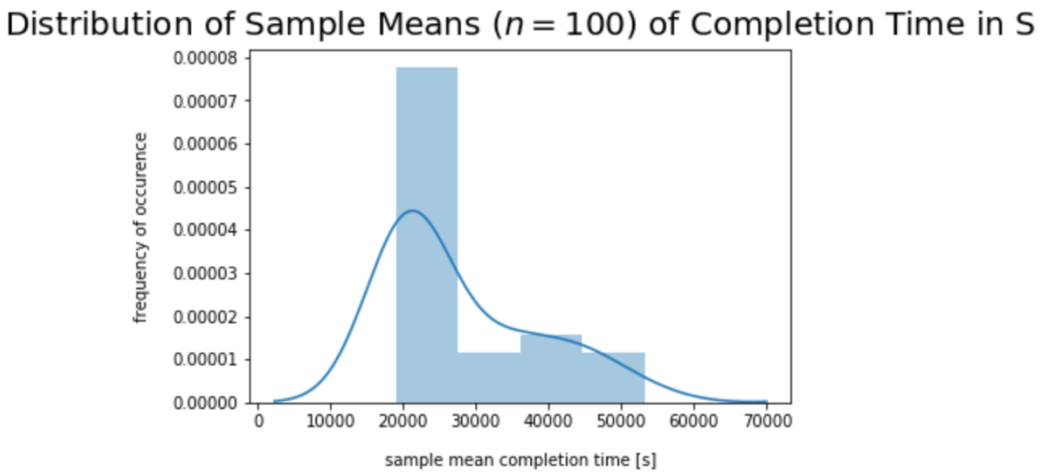

Figure 7: Distribution of Sample Means of Completion Time in S

I conduct an example with the first group (simulation_id=904, algorithm_id=8) to detect the probability of seeing a random sample mean from all sample means less than group 1sample mean time (26242.263465) based on the score z-score FIGURE 8. In addition, I construct for 95

% of lower and upper bound for the first group (simulation_id=904, algorithm_id=8), and plot in Figure 9.

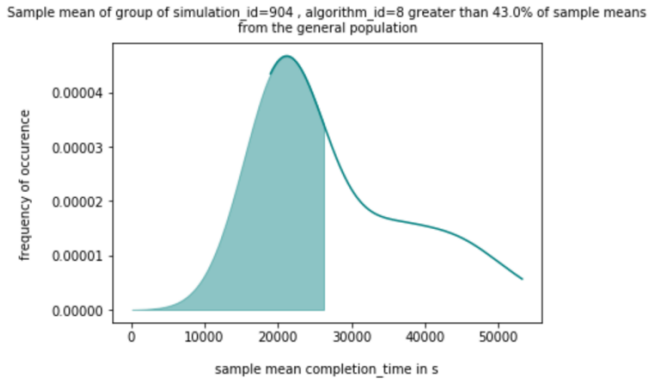

Figure 8: Distribution of Sample Means of Completion Time in S

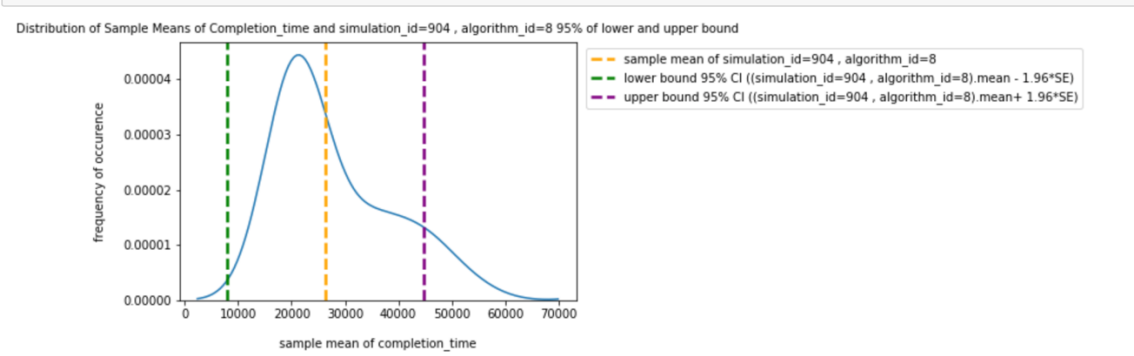

Figure 9: Lower and upper bound of group of simulation_id=904, algorithm_id=8 probability.png

Finally, I calculated of every z-score of each group to detect the probability of seeing a random sample mean from all sample means less than a group sample mean time and constructed 95% confidence interval for all groups. Regard of 95% I might express that I am 95% confident that the value of chosen group is in range of calculated confidence interval. I merged every property in result data frame Figure 10. Furthermore, I visualized all 95% confidence interval for each group in Figure 11. Each numbers represent index of groups which is given in Figure 10.

	simulation_id	algorithm_id	mean_time	std_time	z_scores	95%_confidence_intervall
0	904	8	26242.263465	87942.240552	-0.177572	(7843.80991731372, 44640.71701190995)
1	909	8	27192.605234	87962.901918	-0.076333	(8794.15168715782, 45591.05878175405)
2	914	8	33145.973159	93281.204827	0.557872	(14747.519611801967, 51544.4267063982)
3	952	14	18993.889209	80274.289625	-0.949732	(595.4356618328056, 37392.34275642903)
4	953	14	19332.347864	78762.953849	-0.913677	(933.8943169371378, 37730.801411533364)
5	954	14	22445.019312	81229.450887	-0.582088	(4046.565764571991, 40843.47285916822)
6	1020	18	19308.411174	84799.124694	-0.916226	(909.9576263751514, 37706.86472097138)
7	1022	18	19876.670717	83270.959669	-0.855690	(1478.217170055821, 38275.124264652055)
8	1024	18	22843.440396	85241.572022	-0.539644	(4444.986848932182, 41241.89394352841)
9	1068	19	19102.616134	81420.675156	-0.938150	(704.1625867459188, 37501.06968134215)
10	1069	19	19530.597065	79956.952409	-0.892557	(1132.1435174409307, 37929.05061203716)
11	1070	19	22478.948275	82009.606532	-0.578473	(4080.494728128142, 40877.40182272437)
12	1164	25	43248.974257	125218.224614	1.634131	(24850.5207093415, 61647.42780393773)
13	1165	26	34937.162318	97990.522819	0.748685	(16538.70877047711, 53335.61586507334)
14	1166	27	34390.495417	96395.121577	0.690449	(15992.041869805027, 52788.94896440126)
15	1167	25	47292.943289	129729.245822	2.064930	(28894.48974194117, 65691.3968365374)
16	1168	26	38244.597965	102996.144585	1.101022	(19846.14441807069, 56643.05151266692)
17	1169	27	36705.142686	98203.540971	0.937025	(18306.689138206457, 55103.59623280269)
18	1170	25	53277.846387	129756.006393	2.702495	(34879.39283980218, 71676.29993439841)
19	1171	26	45022.817759	107717.022704	1.823097	(26624.36421163848, 63421.27130623471)
20	1172	27	44308.622195	105172.288745	1.747014	(25910.168647395927, 62707.07574199216)
21	1194	28	20514.358183	91671.736833	-0.787758	(2115.904636201347, 38912.81173079758)
22	1196	28	21038.259475	89741.403107	-0.731948	(2639.8059277360953, 39436.713022332326)
23	1197	29	20080.583842	81779.719817	-0.833968	(1682.1302946382166, 38479.03738923445)
24	1198	28	24496.227765	91701.015226	-0.363575	(6097.774217500068, 42894.6813120963)
25	1199	29	23554.583654	84366.389366	-0.463887	(5156.1301068937755, 41953.037201490006)

Figure 10: Result Dataframe

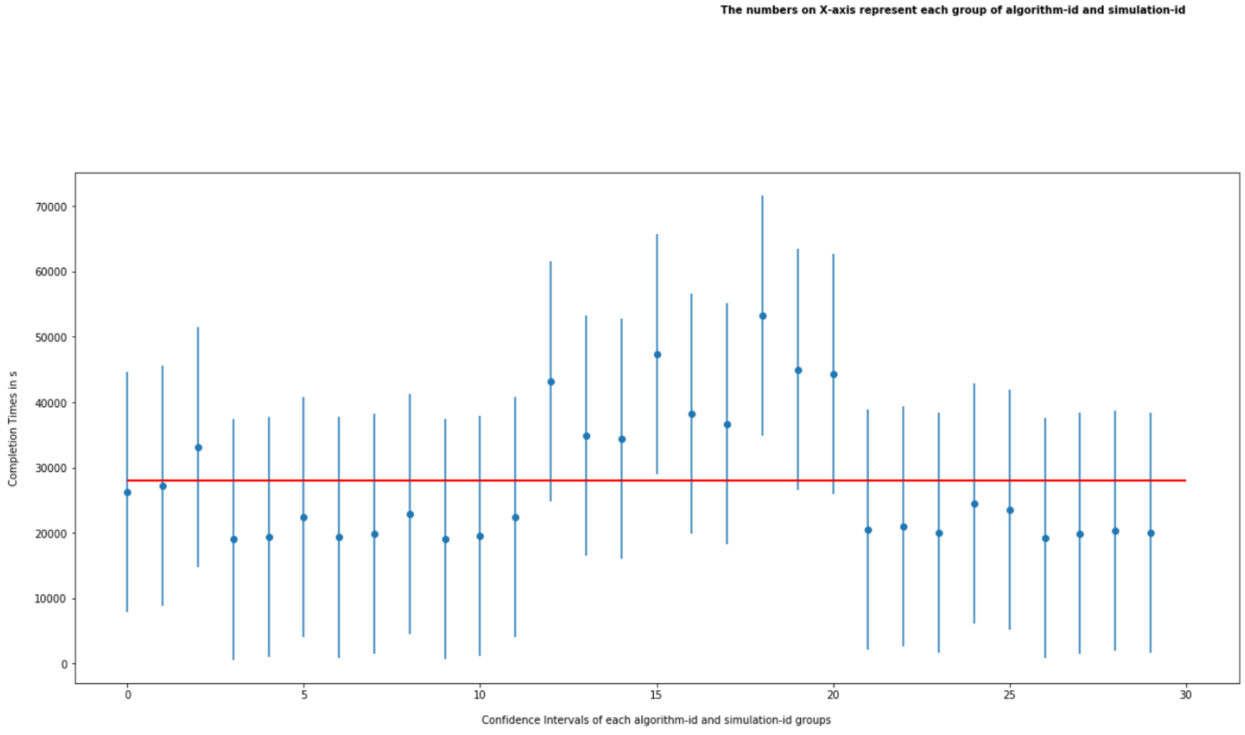

Figure 11: Confidence Intervall