1 INTRODUCTION

The Hyde Housing Group is a very inspirational group that has the strong social mission of providing affordable, well-built housing for people that have been "left behind by the market". They can provide this wonderful housing because they are a not-for-profit that uses income generated by their profit-making and charity subsidiaries that make up the Hyde Group. They also have many practices that make it possible for people who generally would not be able to afford a home in a certain area to do so. One of these practices that we will look at here is there particular approach to dealing with debt that is accrued by their tenants (arrears). When a tenant misses a payment, instead of starting eviction proceedings or charging accruing fees they work with the tenants to deal with the problem in a way that is beneficial to both parties. This could be in the form of waiting a few months till the tenant can provide the full missed amount or by adding a small amount of the arrear on to their rent payments for a number of months. They have a very lenient and fluid process for this. However, due to a government ruling that is set to decrease rent by 1% for the following four years, beginning in 2016, Hyde is looking at a loss of millions of pounds of income. Since this measure will drastically hamper their ability to build affordable homes, we were asked to look into their arrear history answer a few key questions that will provide the Hyde group with knowledge that will allow them to keep up their standards and practices. The questions that we set out to answer where:

1. How has the amount of arrears changed over the past four years? (not including 2017)
2. What anomalies can we find across our data?
3. Are arrears predictable by factors such as demographics, property type, length of tenancy, geography?
4. Is it possible to determine the best time to pursue an arrear so that the arrangement is the most beneficial to both parties?

This section will be followed by various sections, in section 2 we describe the methods followed to research each question with a brief on pre-processing. In section 3 we talk about our results from the methods and models described in previous section. Finally, in section 4 we provided a summary of results and our reflection upon the approach taken to answer the research questions.

2 METHODOLOGY

2.1 Research strategy

The first step of our research strategy was to identify how the arrears are recorded and in what levels within Hyde's accounting system. The arrears are recorded under different tabs within the dataset. It then was needed to create a general model that can determine the factors that affect the change arrears over time, and to make predictions. Because investigating the relationships between different variables are important and Generalized Linear Regression was an appropriate model to apply. Also, the Upper and Lower fence method were used to find outliers in order to identify anomalies.

2.2 Preprocessing

For this project we were given 18 text files totaling 68 Gb that were represented in a star schema containing 8 dimension tables and 2 fact tables. The largest text files were 9Gb each and we had about 15Gb of data for each year. With these huge amounts of data we concluded quickly that we would not be able to process this data on the computers we have available to us. So, we asked the Data Science department and Lancaster University for room on a server to hold our data and run our larger pieces of code. We also decided that a good amount of our programming should happen in python as it deals a bit better with large data sets (we used *numpy* and *pandas* libraries). Once we gained access to the server we loaded our data onto it and installed a virtual environment where we could control the version of python and the packages we wanted to use. Once this was set up we wrote code that would run through the files and perform dimension reduction and some format changes. For this process, we selected columns from the fact tables that we though would be useful for our analysis and we ran a date-time conversion function (*pandas.to_datetime*) on the date column of the data to produce day, month, and year columns. In total process took a file with 58 columns and produced a csv file with 10 columns, including our engineered features, with a total size of only ~ 2Gb. Once we had these files we were able to read them into our local computers much easier and manipulate them. It was during this the process of reshaping and converting the data that we became aware of a serious problem with the data as it was sent to us. All of the text files that we had been sent are corrupted and this resulted in a massive quantity of missing data (Figure 1).

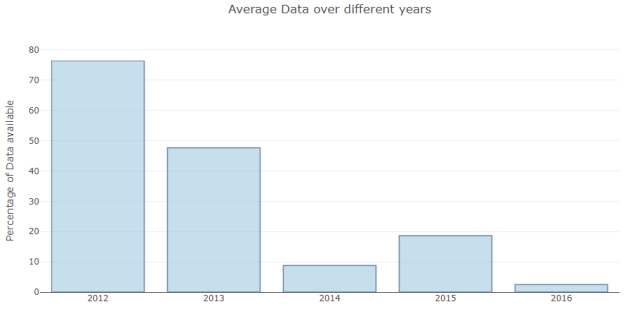

Figure 1: Average data available over different years

Because we were unable to receive new data from the Hyde Group in time we were forced to use these corrupted files the best we could to answer our questions.

2.3 Methods applied for question 1

The first question that we endeavored to answer was "How has the amount of arrears changed over time?". To get the most representative value, we calculated the sum of arrears from *FAB_CURRENT_ARREARS_BALANCE_TO*, and graphed the sum of each month in a monthly plot. This gave us a high resolution look at how the total amount of arrears owed to the company changes day to day. We felt this analysis would be useful in determining trends during the course of a month.

Next we wanted to explore how the arrears balance changed over the course of a year. To this end we took the sum of the daily arrears and calculated the mean across the entire month. This gave us a single value that represented the mean daily balance for each month.

Finally, we investigated the average numbers of weeks that arrears have been due for each month of the year. Suppose a tenant misses a payment, each week after that day till he pays some or the whole amount, the *FAB_NUM_WEEKS_IN_ARREARS* will keep on increasing by 1 as each week passes. For this we calculated the mean of *FAB_NUM_WEEKS_IN_ARREARS* over each month's data that we singled out in the first step. With this we could plot and see how the mean of *FAB_NUM_WEEKS_IN_ARREARS* vary for each month.

2.4 Methods applied for question 2

In an attempt to discover any anomalies in the arrear data we first approached the problem with the intent to look at the distribution of the data, choose a meaningful Z score, and discover the values that would lie outside the interval we create. To do this we first had to remove all missing data from the *FAB_BALANCE_AMOUNT* and make sure that the data we were using followed a Gaussian distribution. As we can see in Figure 2 our data does approximate one half of a Gaussian curve. Once we had confirmed this we applied python functions to find the mean and standard deviation of this dataset. Once we have those we can calculate the Z score using the function:

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

While calculating the terms we mirrored our data to the negative axis in order to deal with a full Normalized Gaussian. After this we were able to apply the above equation to determine the Z score for each value. However, the flaw with the Z score is that it is high susceptible to outliers changing the values of essential parameters. We saw evidence of this during our analysis and that lead us to abandon the Z score test in favor of the "Upper and Lower Fences" (ULF) method.

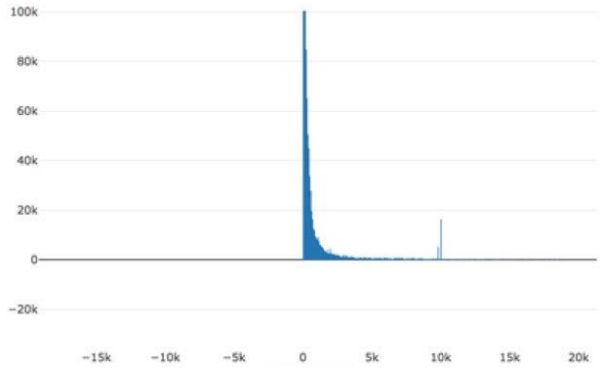

Figure 2. Gaussian distribution curve

The ULF method is a method developed by John Wilder Tukey for use in creating box plots that show outlier data. The method is as follows:

$$Upper\ Fence = Q3 + (1.5 * IQR)$$
$$Lower\ Fence = Q1 - (1.5 * IQR)$$
$$where:$$
$$IQR = Interquartile\ range$$
$$Q3 = Third\ quartile\ value$$
$$Q1 = First\ quartile\ value$$

The 1.5 multiplier applied to the interquartile range is by convention. It was regarded by Dr. Tukey as good in-between value to use because he felt that 1 would be too small and 2 is too large. It's similar to the common practice of using certain values for P-values in statistics.

2.5 Methods applied for question 3

We are interested in factors such as demographics, geography, and length of tenancy in order to discover their effects on arrears. Based on those factors, below columns (Table 1) are selected out of *DM_TENANCY* (871150 rows and 102 columns):

Variable	Variable description
DTE_POSTAL_ADDRESS	Address of the tenant (factor)
Postcode	Postcode of the tenant, extracted from DTE_POSTAL_ADDRESS by using regex '/[A-Z]{1,2}[0-9][0-9A-Z]?\\s?[0-9][A-Z]{2}/' (factor)
DTE_DATE_OF_BIRTH	Birthdate of the tenant with dd/mm/yyyy (factor)
Age	Number of years of the tenant
DTE_SEX	Gender of the tenant (factor)
DTE_ETHINIC_ORIGIN	Ethnic origin of the tenant (factor)
DTE_GEOGRAPHIC_ORIGIN	Country origin of the tenant (factor)
DTE_LANGUAGE	Language spoken by the tenant (factor)
DTE_SEXUALITY	Sexuality of the tenant (factor)
DTE_RELIGION	Religion of the tenant (factor)
DTE_NATIONALITY	Current nationality of the tenant (factor)
DTE_DIM_START_DATE	Tenancy start date (factor)
DTE_DIM_END_DATE	Tenancy end date (factor)
Length of tenancy	Nnumber of weeks (continuous)

Table 1. Description of variables

After that, we merged the filtered tenancy table and the account balance table for 2012 along with other columns below *FAB_BALANCE_DATE*, *FAB_BALANCE_AMOUNT*,

FAB_NUM_WEEKS_IN_ARREARS, and the index of the account balance table, in order to get the necessary dataset. Some specific cleaning tasks are needed. Firstly, the postcode is not provided, so we extracted the postcode from the *DTE_POSTAL_ADDRESS* using the regex mentioned above. Secondly, age of the tenant is not provided so we counted the year difference between today's date and the date of birth of tenant. Thirdly, the table had different time format, some are "dd/mm/yyyy" and some are "dd/mm/yyyy HH:MM:SS" so we treated the date as a string and take the first 10 characters so all the date will have same format which is "dd/mm/yyyy".

In order to understand the relationship of among factors, we used generalized linear model, which can be written as:

$$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \ldots$$

with

$$y_i = \mathcal{N}(\mu_i, \sigma^2)$$

The equation shows that each of the response variable y is normally distributed with mean and variance 2. βi is the coefficient of variable i and xi is the i-th variable. The model had "FAB_BALANCE_AMOUNT" as the response variable and *tenancy_post_code*, *tenant_age*, *tenancy_length_week*, *DTE_SEX*, *DTE_ETHNIC_ORIGIN*, *DTE_GEOGRAPHIC_ORIGIN*, *DTE_LANGUAGE*, *DTE_SEXUALITY*, *DTE_RELIGION*, and *DTE_NATIONALITY* as the predictors. After that, we did analysis of variance with type 2 sum of square to find significant variables with p-value lower than 0.05. We used anova with type 2 sum of square because we assumed that each of the variables had no interaction to the other variables and type 2 sum of square is powerful for such condition. The result of the anova shows that there are two significant variables, *tenancy_post_code* and *tenancy_length_week*. However, the slope of variable *tenancy_length_week* is really small (0.71) and close to a constant, so we concluded that this variable was not a good predictor variable to predict the balance amount because the balance amount is obviously not a constant.

2.6 Methods applied for question 4

The forth question that we focused to answer was What is the optimal time to pursue arrears? To answer this question, we started with the preprocessed data and cut it to get the features we were interested in:

Variable	Variable description
FAB_RENT_ACCOUNT_KEY	account numbers used to identify individual renters
FAB_TENANCY_KEY	used to identify homes
FAB_IN_ARREARS	Z=Zero Arrear N=Negative arrear Y=Positive arrear
FAB_BALANCE_DATE	date term

Table 2. Description of variables

We chose these features because we wanted to populate a list of values tell us how long people took to pay off their debt. To do this we used an algorithm that is very similar to below.

```
Get list of unique renters key
    For value in renters key:
        Subset data where renters key == value
        For i in len(subset)
            If subset[i] changes from a Z or N to a Y:
                Log start time
            If subset[i] changes from Y to Z or N:
                Log end time
```

This algorithm returned two dates that we could use in our analysis to obtain a better understanding of the amount of time that passed than if we had used the *FAB_NUM_WEEKS_IN_ARREARS* provided in the dataset which only deals with weeks. The downside of this algorithm is that it was very computationally expensive and it took well over 11hrs to run on one year of data. We tried several techniques in an attempt to bring down the time including: parallel processing, writing parts of the code in Python, and using the *map* function to drop parts of the process into C but these attempts either failed or did not decrease our run time significantly. For this reason, we were only able to run our algorithm over 100% of the data in the 2012 dataset.

After running our code in Python to gain the values we wanted we transferred the data to R so we could analyze it. We engineered a new feature called *TOTAL_DAYS* that was built from the difference between the end time and the start time that we recorded. We then grouped our data by this *TOTAL_DAYS* feature and counted the number of rental accounts that paid for value for day. We then took this process a step further and ran the analysis for every month in a year to see if there are any similarities or differences that we could detect month to month.

3 RESULTS AND DISCUSSION

3.1 Results for question 1

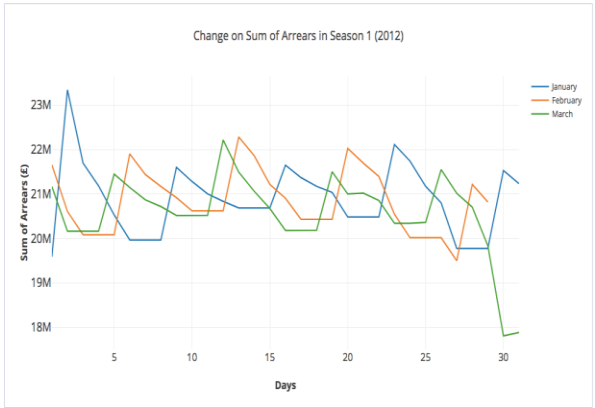

Figure 3. Line graph showing change in sum of arrears- 2012

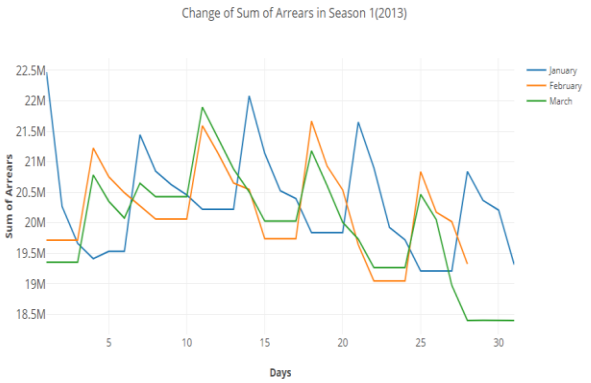

Figure 4. Line graph showing change in sum of arrears-2013

We can see from Figure 3 and Figure 4 how the sum of arrears changes on day to day basis for the 1st quarter. As we can see from both the Figures that there is a spike in the early days of January, this means that there is a general increase in the arrear balance at the beginning of January. This spike in early January can be explained by people having difficulty in paying their bills after spending extra money during the holiday season from previous year.

Moreover, we can see recurring spikes during all of these months in weekly scale, where these spikes are roughly a week apart in most cases. This can be explained by the UK rental payment system where it generates weekly updates on rents for tenants. For Example, in Figure 3, the first spike in January happens on Monday, the 2nd, and as the sum of arrears decreases during the week, it keeps consistent on the weekend (7th & 8th), the graph line increases back to a spike subsequently on the next Monday which is the day Hyde Housing updates rents for next weekly cycle. This trend would apply to all of the weeks for both 2012 and 2013.

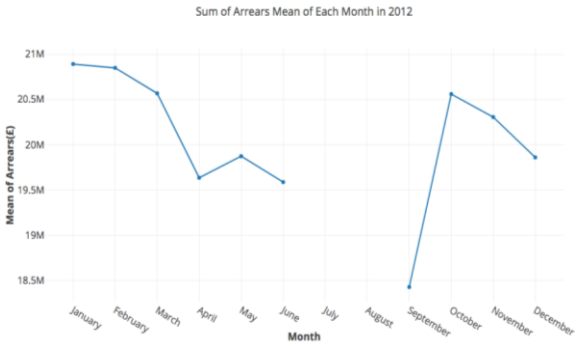

Figure 5. Line graph showing mean arrear for months-2012

Looking at Figure 5 we can see how the mean amount of debt owed every month changes over the year. Unfortunately, we do have a gap for 2 months of the year so our trend is incomplete, however we can still see that the amount due does decrease about 1M from 20.7M to 19.7M per month for the 2nd quarter (April, May, June) of 2012 as we move from the post-holiday season into summer. This decrease is followed by a spike as we go back into the fall season in October.

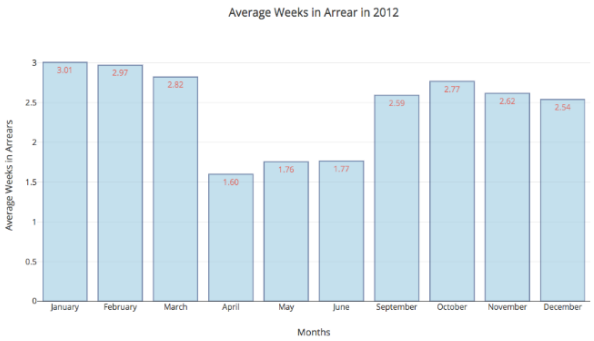

Figure 6. Histogram showing Average weeks in arrear

Figure 6 shows the mean amount of time that people wait before paying off their debts, where the average weeks that people have arrears in their account is about 3 or more, except in 2nd quarter, and this provides an insight on when would be a good time to collect arrears for later analysis. As of in 2nd quarter, it takes less time for people to pay their arrears, where arrears are generally paid off within 1.5 weeks as half of the time than in other months.

3.2 Results for question 2

Year	Lower quartile (Q1)	Upper quartile (Q3)	Inter-quartile Range (IQR)	Upper fence	Lower fence	Censored client amount of outlier with percentage
2012	£118.28	£ 558.63	£ 440.35	£1219.16	£0	993886 (2.74%)
2013	£117.2	£569.67	£452.47	£1248.36	£0	607312 (2.61%)
2014	£112.75	£591.83	£479.08	£1310.45	£0	98081 (2.39%)
2015	£132.61	£624.28	£491.67	£1361.79	£0	205917 (2.17%)
2016	£124.93	£630.95	£506.02	£1389.98	£0	23644 (2.25%)

Table 3. Upper and Lower fence method for every year

Table 3 above shows the results after implementing Upper and Lower fence method for all years' data. According to Hyde's accounting system, arrears are recorded as a positive amount under *FAB_CURRENT_ARREARS_BALANCE_TO* for tenants who did not pay rents in time, so the lower fence is set to £0. The analysis of 2012 shows that 2.74% out of all date falls outside of the range of [lower fence, upper fence], and are marked as outliers. This percentage falls as the year continues. Another interesting finding from above table is that the Inter-Quartile Range gradually increases from £440.35 in 2012 to £506.02 in 2016, which means that the median arrear amounts increases implying the amount of rents increases as year increases.

3.3 Results for question 3

To predict *FAB_BALANCE_AMOUNT* by using postcode, we separated the table into two after we randomized the order of the table. 80% of the data were designed for training step and the rest for testing. We used the training data to fit the model, and testing data to predict by calculating the mean absolute error between the real balance amount and the predicted one. The mean absolute error of this dataset was 204.35. To achieve lower error, we removed the outlier that are located outside this range (q1-1.5*iqr, q3+1.5*iqr), with q1 being the first quartile, q3 the third quartile, and iqr is the interquartile range which is the difference between q3 and q1. This outlier removal method was the same as we did in question 2. After the outlier being removed, we predicted the *FAB_BALANCE_AMOUNT* by using postcode again, and got the mean absolute error 119.46.
Next, similar prediction can be implemented by using only positive amount balance including 0, because arrears is a positive amount balance and 0 means that the arrears has been paid off. The result shows that the mean absolute error for the prediction is 199.22, and this number decreases to 124.07 after outlier removal being applied.

3.4 Results for question 4

After the analysis for method in section 2.6 we were able to produce the graph in figure 7. In Figure 7, it shows a bar chart of the number of payments by range of days. We can see from the graph that the majority of arrears are paid off either one week or four weeks after accruing the debt. Thus, the majority amount of payments is made during the first four weeks compared to all range of days.

After the first four weeks, in range of days ((31,61], (61,91], (91,121] and (121,361] days) the number of payment varies between 349 and 885. It can be observed that the highest number of payments are made between 92 and 121 days I.e. 885, in other words, there are more payments during the fourth month after getting into debt. In addition, the lowest number of payments are made during the third month (61,91] after getting into arrear, it is almost half of the number of payments during the range (91,121]. Finally, the total number of payment after four months of getting into arrear (121,361] is 537.

We can see in Figure 8, a bar chart of the number of payments by month. In this Figure, it can be observed that September is the month with less number of payments and there are no payments for July and August, but the reason of this is there are missing data in those months, so we excluded those months in the analysis. In the other months, the number of payments is between 2179 and 3327. The highest number of payments is in January (3327) and October (3173). And the lowest number of payments are in April (2179). The behavior in May, June, and November is very similar, around 2500 number of payments.

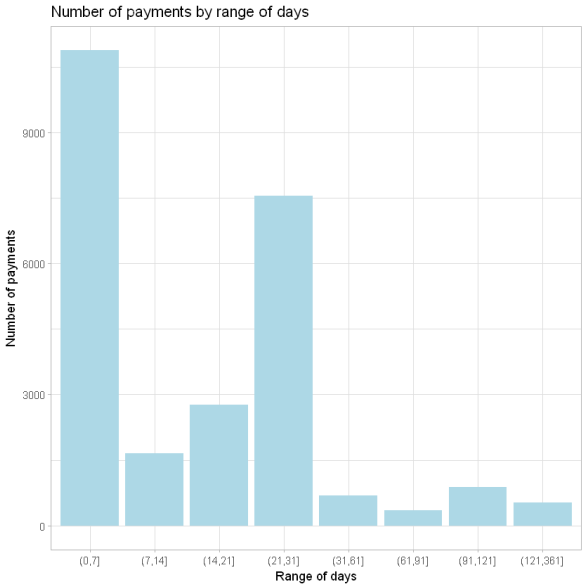

Figure 7. Bar chart of the number of payments by range of days

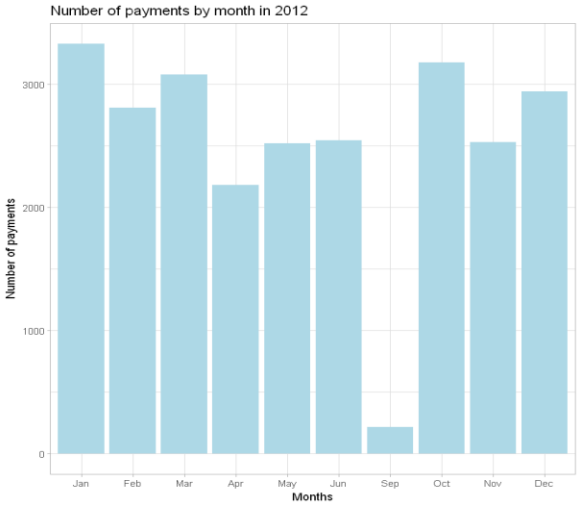

Figure 8. Bar chart of the number of payments by month

3.5 Bias and Validity

It is worth mentioning that the modeling construction was not based on a specific set of variables verified by Hyde Housing, which assumes that our data selection from 871150 columns are comprehensive to Hyde's overall objective. Further still, removing the rows with NA values for demographics while solving question 3, may also have let some bias sink into our research. It should also be noted here that our model for this question can be tweaked so as to get the arrears prediction more accurately for any year as the mean absolute error that we got after removing the outliers is considerably small.

4 CONCLUSIONS

This research investigated the Hyde Housing's arrears condition in order to enhance Hyde's income condition based on the data provided by the Hut Group. In addition, it determined whether theses arrears can be identified by factors such as demographics, property types, length of tenancy, geography, to gain better understanding on the best time to pursue an arrear based on different background each tenant has. Following a thorough data exploration, preprocessing, and cleaning using Upper-Lower-Fence Method, Generalized Linear Model was used to construct the final model based on its variables. The outcome of this study indicated that tenants from Hyde Housing more likely to pay arrears in quarter two, while the same time paying arrears about 2 weeks faster in than keeping arrears up to 4 weeks in other months in 2012. Aligning with this finding, the optimal time to pursue arrears is within 30-120 days. Additionally January or October maybe the best time to purse debts due to the highest number of payments done in those months. Knowing that people generally pay arrears within one month is beneficial for Hyde to construct policies, and build relationship between renters and tenants in order to optimize the amount of return on arrears and to avoid further disputes. The amount of arrears can also be predicted by postcodes of the tenants, where the arrears of residence could affect the price of housing and rental price, where if information can be used to by marketing department to design promotions by areas.

However, there were some emergent limitations encountered. Despite that none of our data set is completely with 12 months, missing values from demographics can be better estimated using Lean++.MF method instead of removing the variables directly, without time constrains [1]. As of further improvements, decision tree could be used instead of ANOVA model for better feature selection.

REFERENCES

[1] Polikar, R. et al. (2010) 'Learn++.MF: A random subspace approach for the missing feature problem', Pattern Recognition. Elsevier, 43(11), pp. 3817–3832. doi: 10.1016/j.patcog.2010.05.028.