Report created by - Siddhesh Bhande

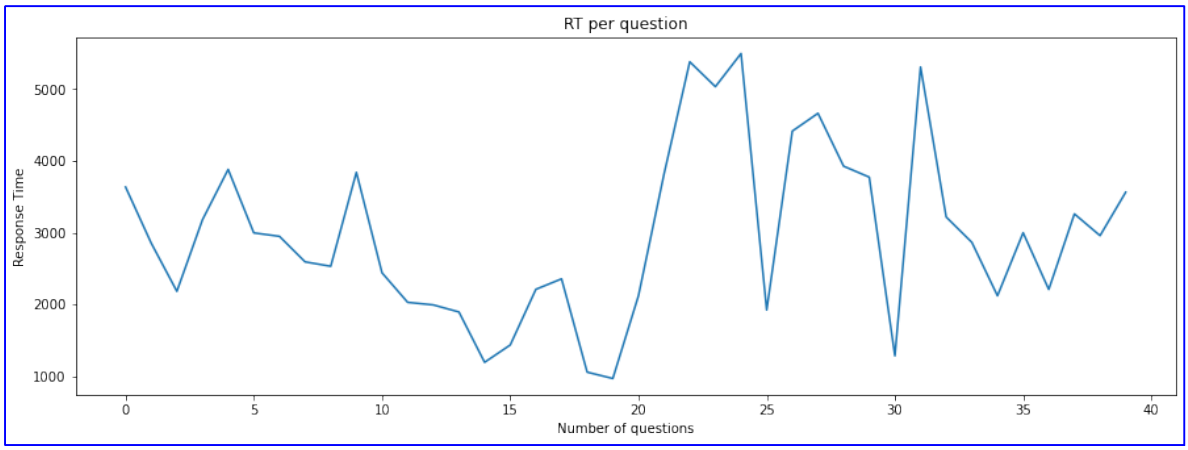
Step 1 - Exploratory Data

	Time	RT	Difficulty	isCorrect
0	1598.902	3635.104895	1	1
1	23814.206	2853.791475	2	1
2	39763.242	2182.772636	2	1
3	61978.531	3180.494785	2	0
4	79677.448	3880.589724	2	1

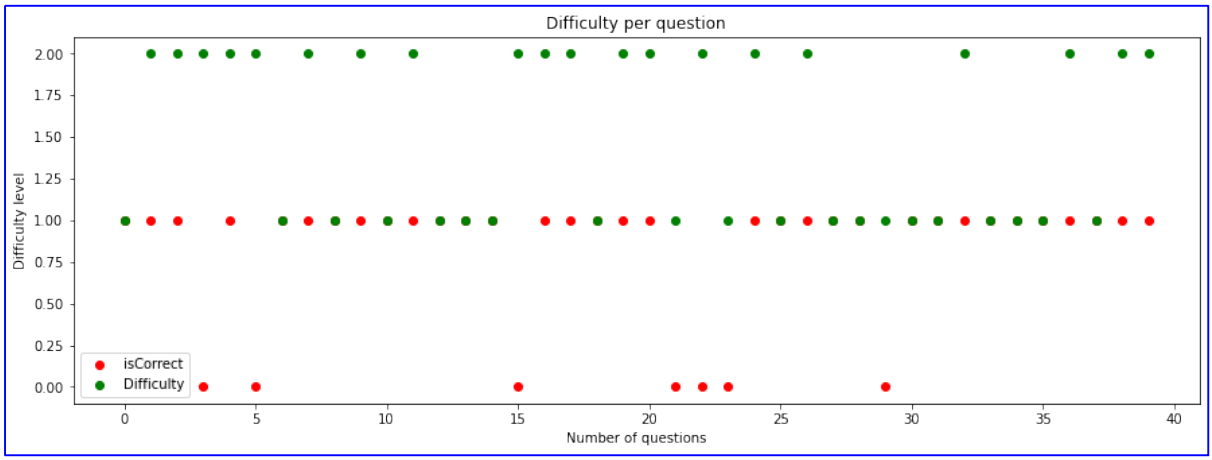
Analysis

1. Let us check for the behavioral data for subject 1

Lets try to visualize the data



Observations:
Here we can see that the response time for question 15-20 is very less and the response time for questions 20-25 is high. We should check the dificulty level of questions between this timeline.



Here we can see for questions 20-25 he got the maximum incorrect reponses when the difficulty level was less and more time was taken

2. Lets try to check the distribution among pupil data for subject 1

	Time	Pupil Diameter
count	189518.000000	189518.000000
mean	379038.000000	4.624206
std	218837.113982	0.629241
min	4.000000	2.844679
25%	189521.000000	4.195024
50%	379038.000000	4.642508
75%	568555.000000	5.075196
max	758072.000000	6.275692

3. First lets check if we have any missing values in the entire data, missing values are usually given by NA.

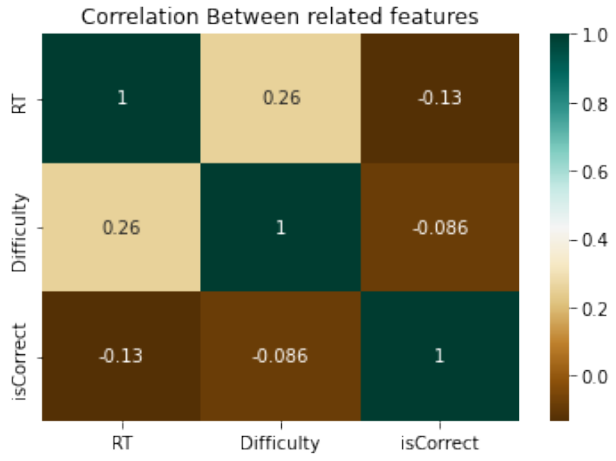
```
#checking if we have any missing value

merged_data.isnull().sum()
```

```
Time           0
RT            72
Difficulty     0
isCorrect      0
dtype: int64
```

```
merged_data = merged_data.dropna()
```

Here the data had 72 rows with NA values. I have used the dropna function to remove all the rows with missing values

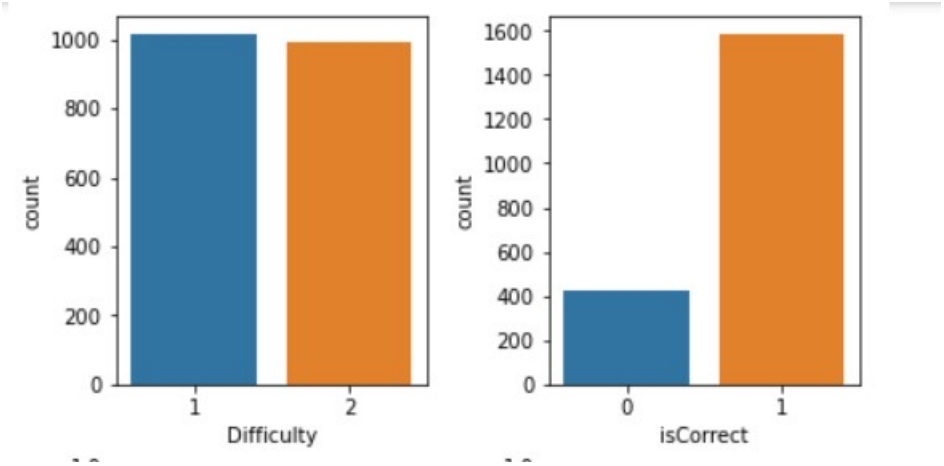
4. Lets try to find if there is any correlation present between the features which can be dependent.
I did this using a heatmap.



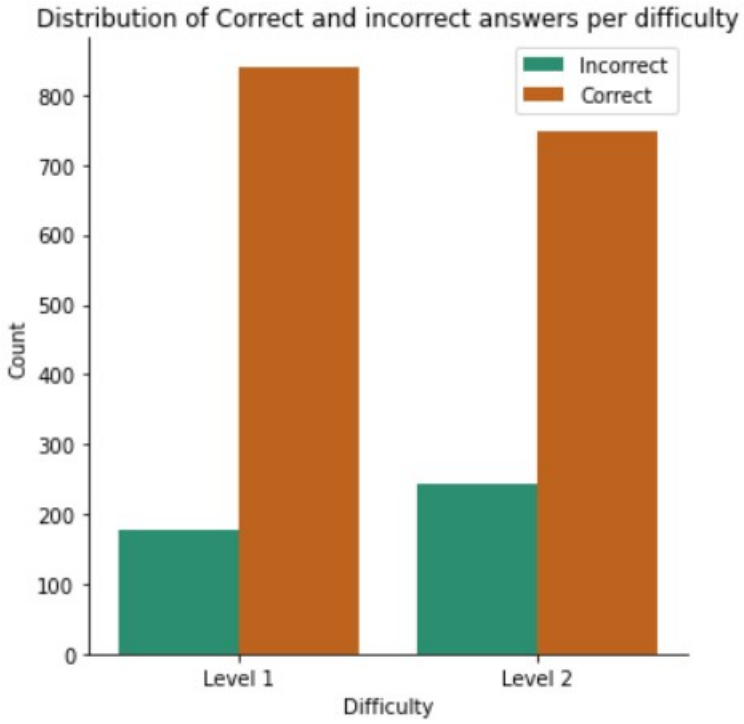
Observations -

Here we can see that response time and difficulty has a positive correlation, which says that if difficulty is more pupil takes more time to reply Also Difficulty and isCorrect has negative correlation, which says if the question is more difficult then value isCorrect is less.

5. For a machine learning model to work correctly, the data should be identically distributed. I tried to check if the given data is identically distributed for the categorical features Difficulty and isCorrect



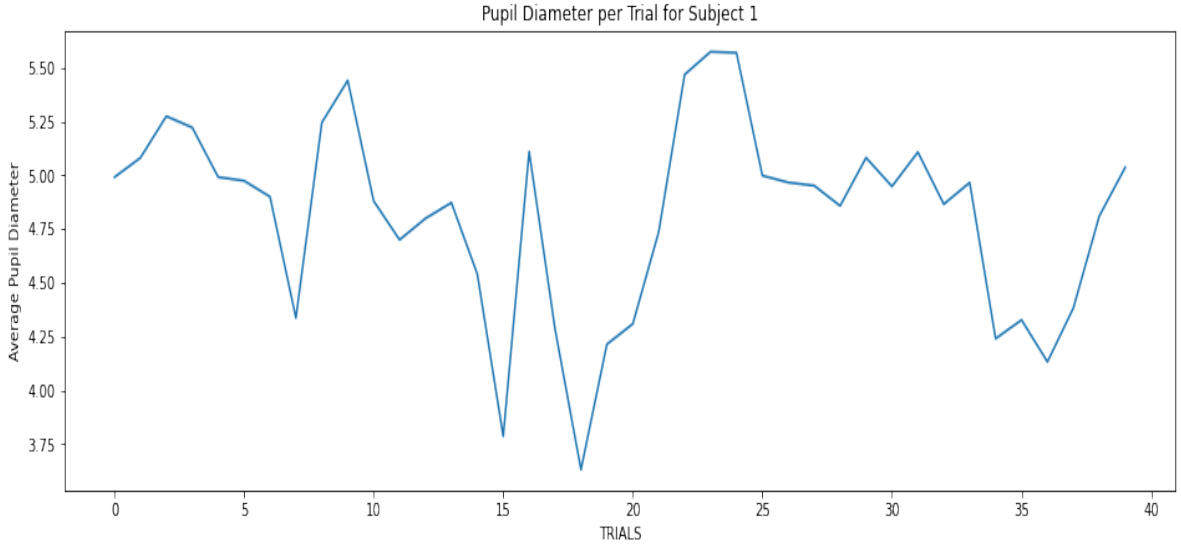
Data Distribution



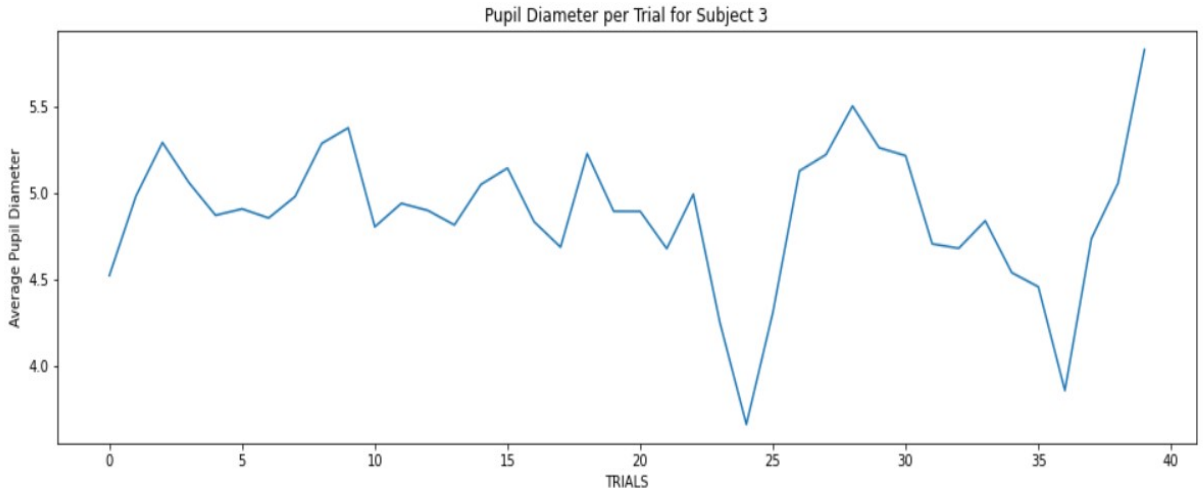
Observations - Number of rows in Difficulty are properly distributed, however the number of Correct answered trials is far greater than the number of incorrect answered. This may cause some issue while training the model.
Resampling techniques can be used in such a scenario.

Step 2: Examination of the pupil responses

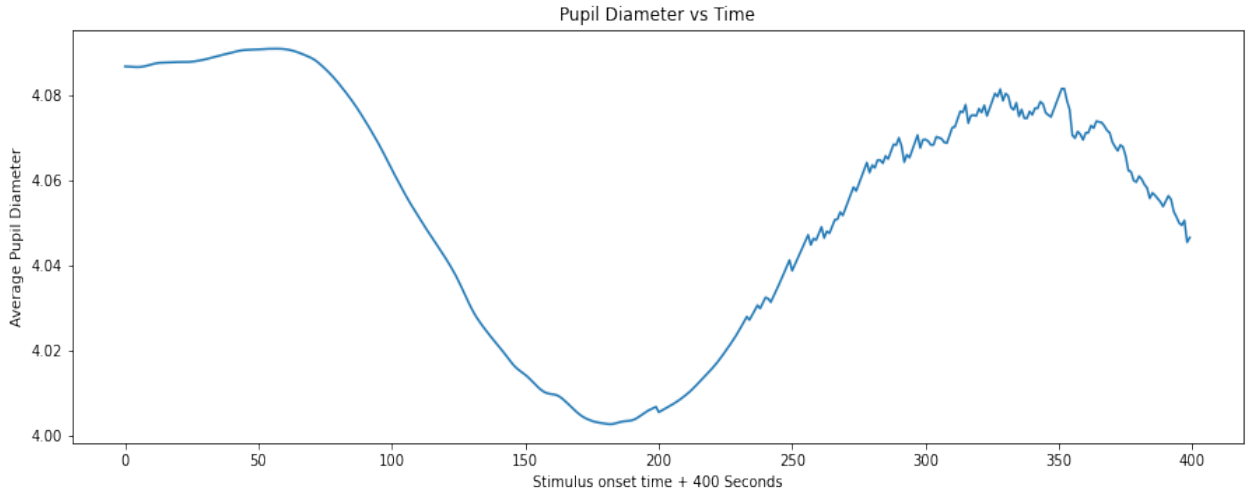
1. Here we have plotted average Pupil diameter per trial for subject 1. We can see that for trials 20 - 25 the average pupil size is larger, we have found in previous distribution step that subject took the maximum time to answer this questions as compared to others.



2. Here we have plotted average Pupil diameter per trial for subject 3. We can see that some of the trials have similar slopes as subject 1.

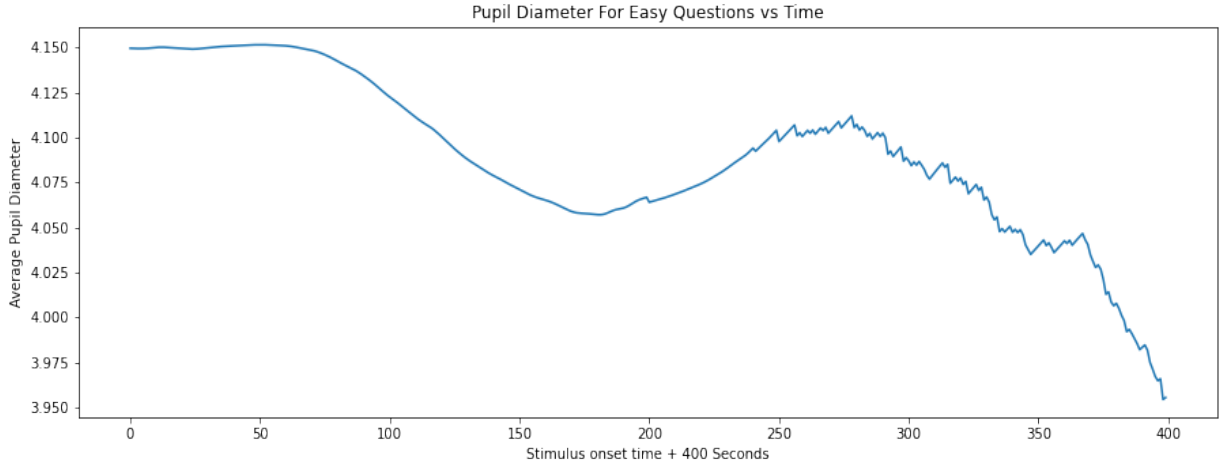


3. Next I have plotted the average pupil size, for each trial, between the period from start onset till start onset + response time across all the subjects.



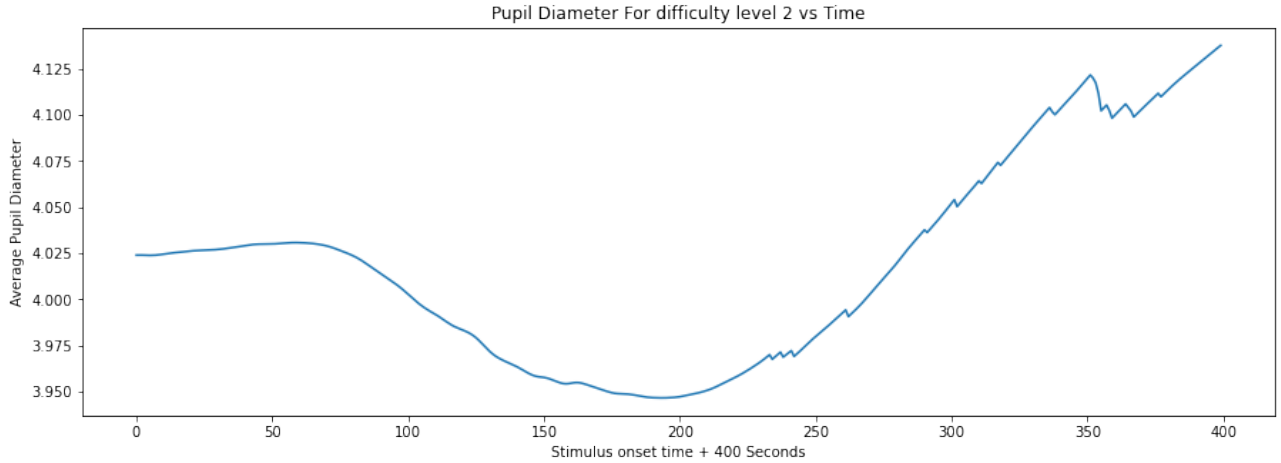
Observations - Here we can see that the pupil size is more during the start of trial, this can be the reaction when question is first visible. After some time between 150-250 seconds the pupil size is reduced. And before the response is made the pupil size is increasing gradually with time.

4. Next I have plotted the average pupil size, for each trial, where the difficulty level is 1 across all the subjects.



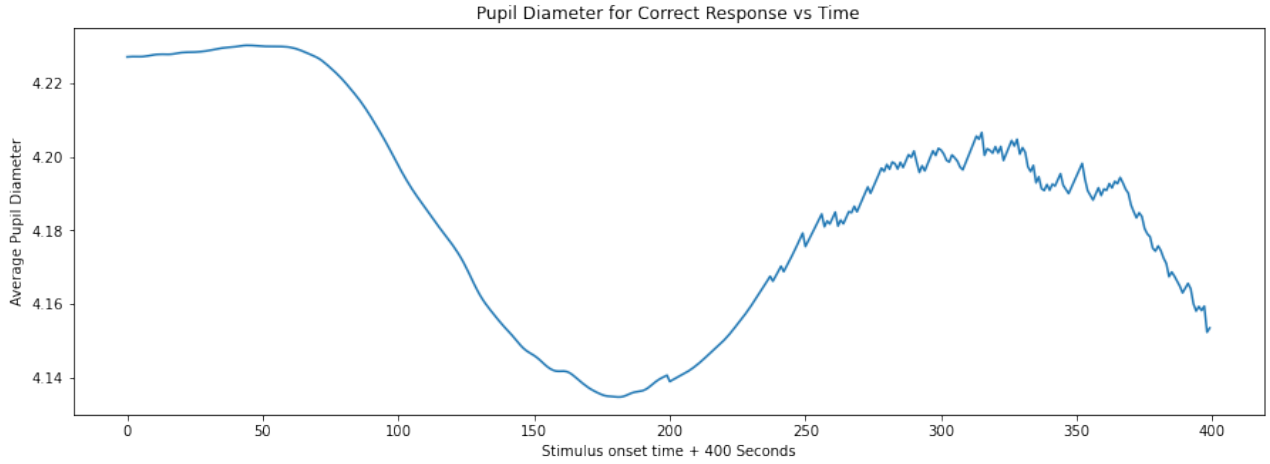
Observations - Here we can see that the pupil size is more during the start of trial, this can be the reaction when question is first visible. But in the end the pupil size is decreased. I think this can be because the subject found the correct answer is now feeling relaxed.

5. Next I have plotted the average pupil size, for each trial, where the difficulty level is 2 across all the subjects.



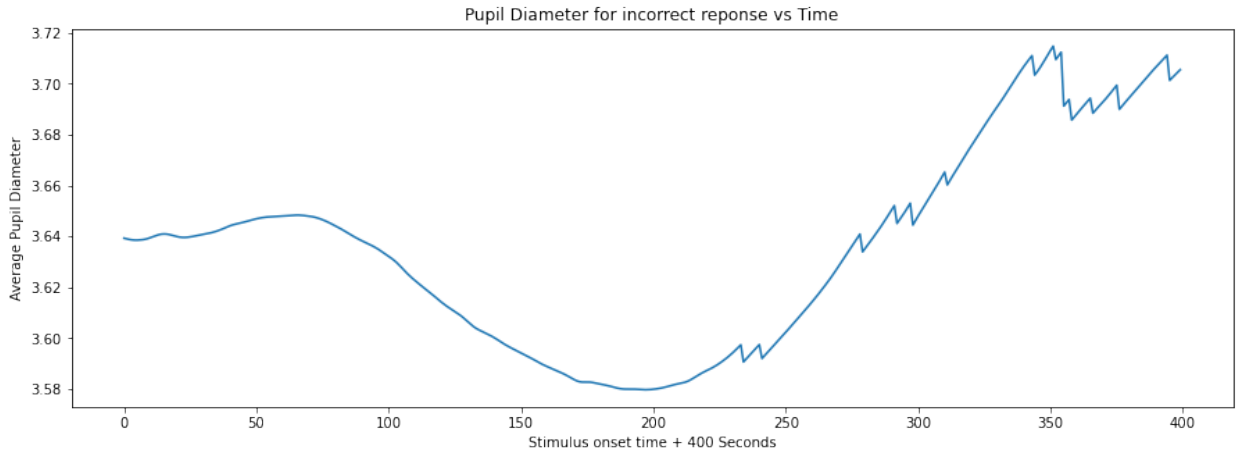
Observations - Here we can see that the pupil size is more during the start of trial, this can be the reaction when question is first visible. But in contrast to above plot, in the end the pupil size is increased. I think this can be because the subject found that the difficulty level is more answer is now feeling anxious.

6. Next I have plotted the average pupil size, for each trial, where the correct responses were made across all the subjects.



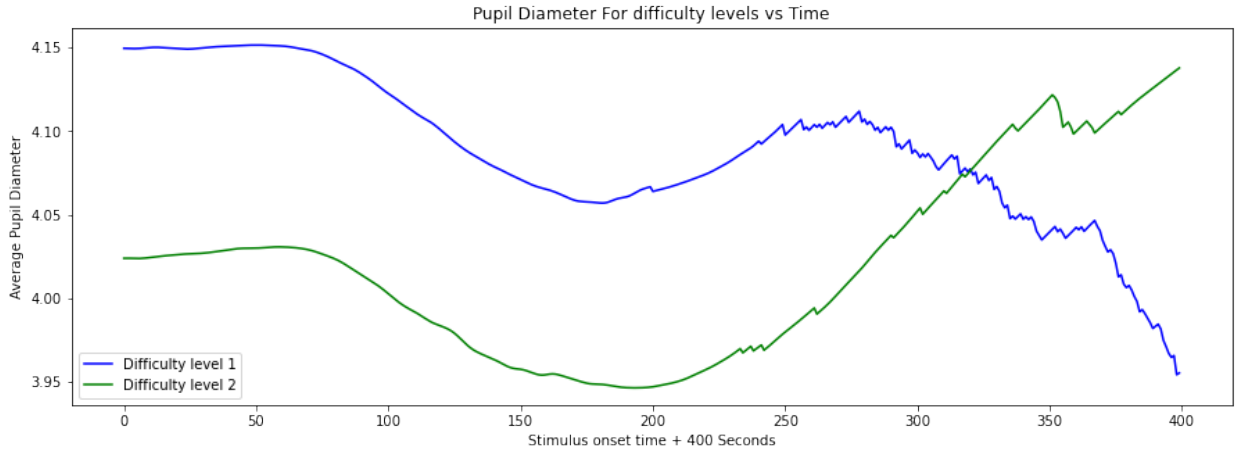
Observations - Here we can see that the pupil size is more during the start of trial, this can be the reaction when question is first visible. But in the end the pupil size is decreased. This same pattern is observed for questions with difficulty level 1.

7. Next I have plotted the average pupil size, for each trial, where the incorrect responses were made.

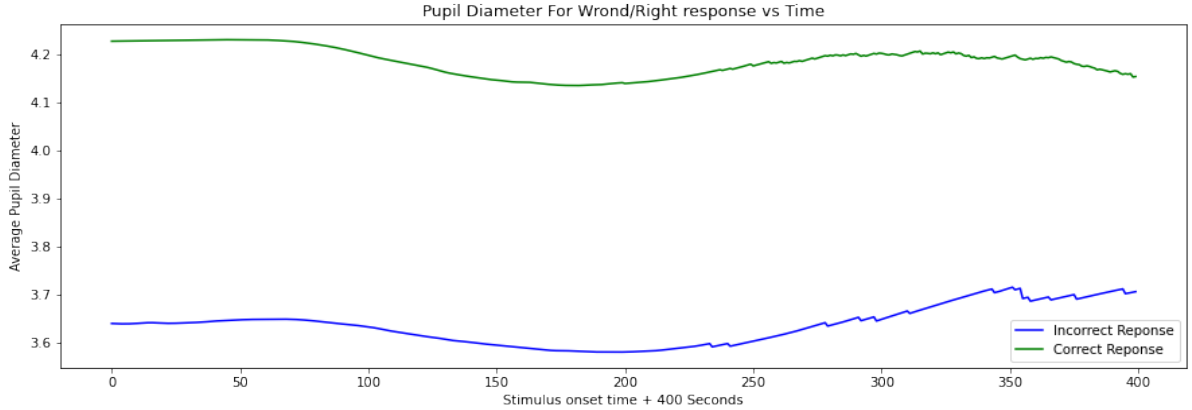


Observations - Please note that we found the number of incorrect responses were less than the number of correct responses. This can be the reason that the average pupil size is reduced for incorrect responses. Here we can see that the pupil size is less than mean during the start of trial. But in the end the pupil size is increased. I think this can be because the subject was not able to reach a correct answer. The pattern is same as for the questions with difficulty level 2.

8. Plotted the average of pupil diameter for easy/hard questions.



9. Plotted the average of pupil diameter for wrong/right responses.



Step 3

So far -
1. We have pupil responses per trial for all the subjects concatenated.
2. We have response time required for each stimulus, its dificulty and correctness

Algorithm to create a dataframe that can be used for prediction -
1. Get the peak pupil response and create new dataframe with only one column
2. Get the index(column name) of peak pupil to get the latency
3. Add RT, difficulty and isCorrect as new columns to this dataset
4. This dataset will be used for training our model

Created Dataframe -

```
newData.head()
```

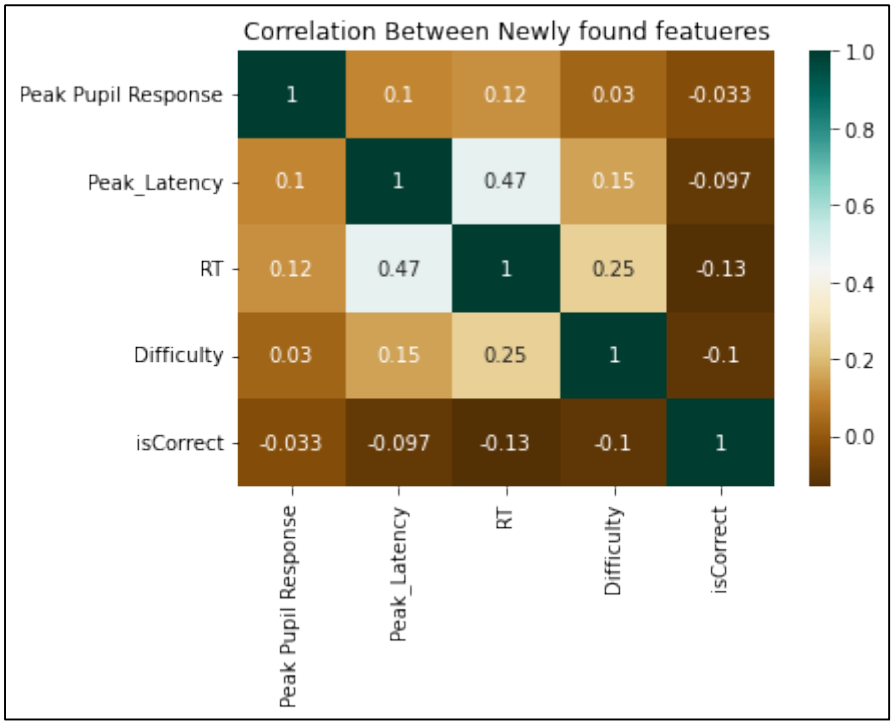
	Peak Pupil Response	Peak_Latency	RT	Difficulty	isCorrect
0	5.991895	3636.0	3635.104895	1	1
1	5.815964	2852.0	2853.791475	2	1
2	5.563529	1552.0	2182.772636	2	1
3	5.691393	3180.0	3180.494785	2	0
4	5.597188	3880.0	3880.589724	2	1

Training of Model -

1. Statistical analysis of data using the describe method to check mean and standard deviation

	Peak Pupil Response	Peak_Latency	RT	Difficulty	isCorrect
count	1969.000000	1969.000000	1969.000000	2040.000000	2040.000000
mean	4.832045	2667.226003	4021.613469	1.500000	0.761275
std	0.900800	1207.235164	1635.687459	0.500123	0.426409
min	3.037058	4.000000	797.361135	1.000000	0.000000
25%	4.197417	1944.000000	2733.876944	1.000000	1.000000
50%	4.813116	2996.000000	3894.249678	1.500000	1.000000
75%	5.343777	3684.000000	5258.509636	2.000000	1.000000
max	7.661647	3964.000000	7992.403507	2.000000	1.000000

2. I tried to find the correlation to check understand the new features

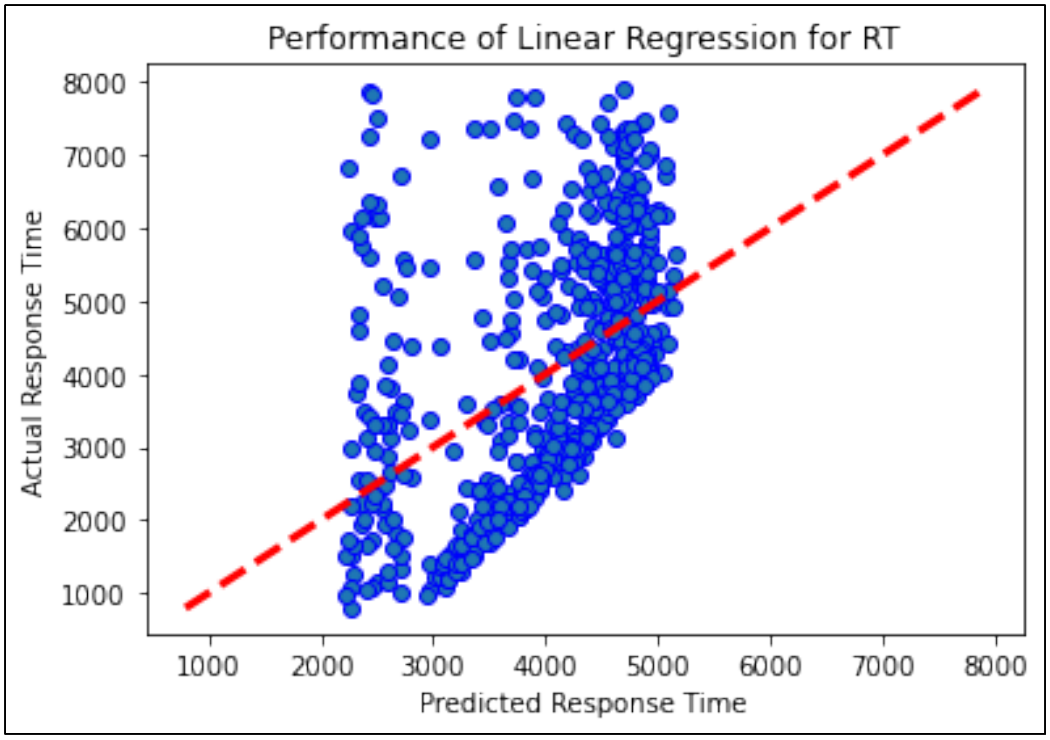


Here you can see some correlation between Response time and peak latency, this can be used in step 5 as well. Also some correlation can be observed between the difficulty and Response time.

3. **Training of model for Response time using first 2 features**

Here as RT is a continuous variable I used linear regression for training. Here are the training results.

Mean absolute error = 1204.41
Mean squared error = 2126294.62
Median absolute error = 1103.34
Explain variance score = 0.24
R2 score = 0.24



Step 4:

I tried to train and test the model for difficulty using the same set of features as described above.

1. I split the data into train and test. I used logistic regression to train the model for predicting the difficulty class which Is a categorical variable.
I got the training accuracy of 60% and testing accuracy of 57%.

I tried to train and test the model for correctness using the same set of features as described above.

2. I split the data into train and test. I used logistic regression to train the model for predicting the isCorrect class which Is a categorical variable.
I got the training accuracy of 79% and testing accuracy of 78%.

Step 5:

Here first I created a test dataframe using the files provided in test folder.

	Peak Pupil Response	Peak_Latency
0	5.991895	3636.0
1	5.815964	2852.0
2	5.563529	1552.0
3	5.691393	3180.0
4	5.597188	3880.0

Predict the Response time based on linear regression model trained in step 3

	Peak Pupil Response	Peak_Latency	Predicted Onset Time
0	5.991895	3636.0	4772.045307
1	5.815964	2852.0	4269.512475
2	5.563529	1552.0	3440.877898
3	5.691393	3180.0	4456.315691
4	5.597188	3880.0	4875.283585
...	...	...	...
2035	6.932581	3952.0	5077.473786
2036	6.994491	3232.0	4642.397170
2037	6.438678	1584.0	3564.052704
2038	6.726127	3964.0	5060.427732
2039	6.690952	3428.0	4726.925510

This are the test results.