Gramener - Product Team - Submission

National Achievement Survey

Problem Statement-

Given Questions-

1. What influences students performance the most?
2. How do boys and girls perform across states?
3. Do students from South Indian states really excel at Math and Science?

Additional Analysis-

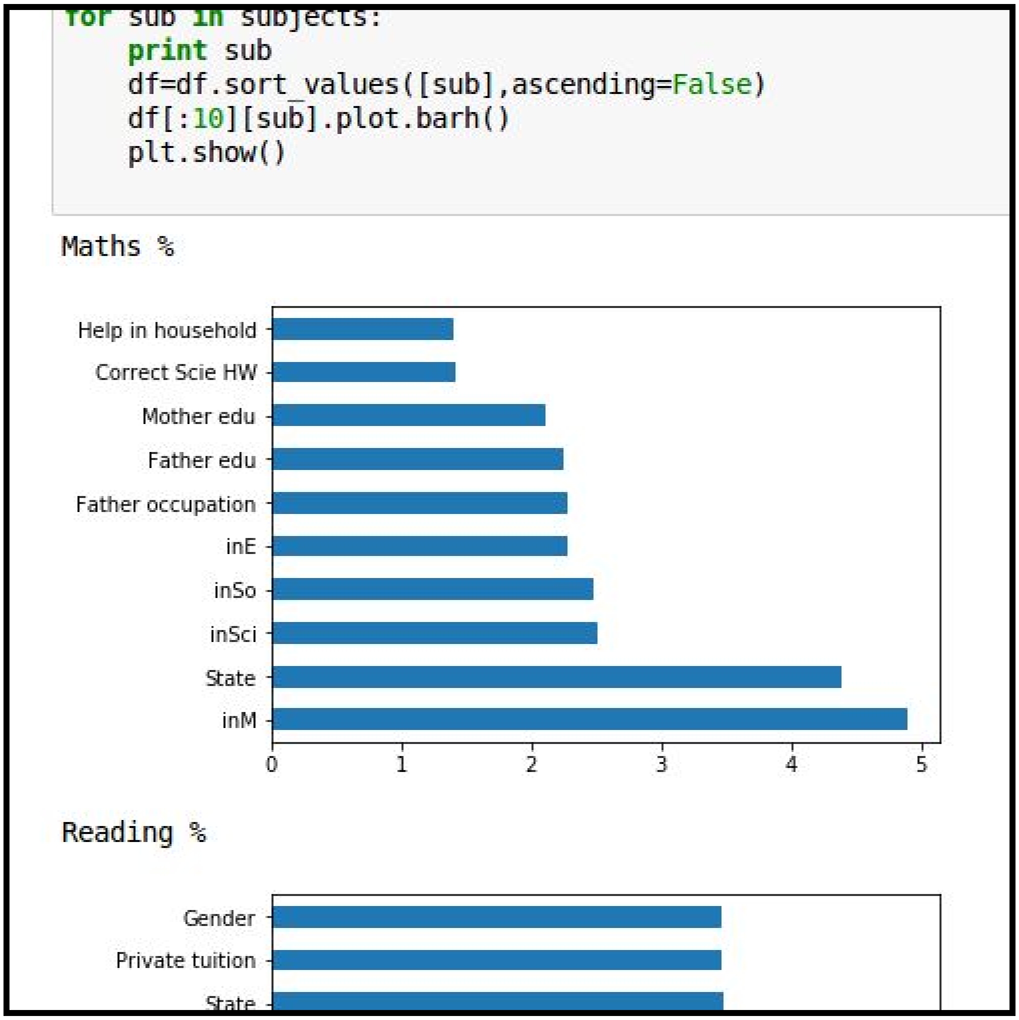
1. Do students who watch TV and use computer score less than those students who do not?
2. Classification model

Question 1. What influences students performance the most?

Approach for identifying factors that influence performance most

- Firstly features like ***'Maths is Difficult', 'Solve maths problems', 'Answer English Aloud'*** etc. were combined to form new features ***'inM'(interest in Maths), 'inE'(interest in English), 'inSci'(interest in Science)*** and ***'inSo'(interest in Social Science)*** as these features are important but didn't had much impact individually, but collectively they turned out to be very important features.
- Each factor consisted of various categories. So for each category median(since it can handle outliers) of the marks was calculated.
- After calculating median for each category, standard deviation was calculated which tells us how spread out these values are i.e how much they differ from each other. For example-
  Factor is '**Mother Education**' and subject is '**Maths**'
  Median of marks for each category in 'Mother Education'
  Category 1 - 29.0
  Category 2 - 20.0
  Category 3 - 47.0
  And for factor '**Age**' and subject '**Maths**'
  Category 1 - 29.0
  Category 2 - 27.0
  Category 3 - 28.0
  We can see above that 'Mother Edu' will have **greater standard deviation** and the median values for each category are spread out and **thus they have more impact.**
- For each subject, factors were arranged in decreasing order and created a plot showing TOP 10 factors.

Screenshots-



```
for sub in subjects:
    print sub
    df=df.sort_values([sub],ascending=False)
    df[:10][sub].plot.barh()
    plt.show()
```

```
Maths %
```

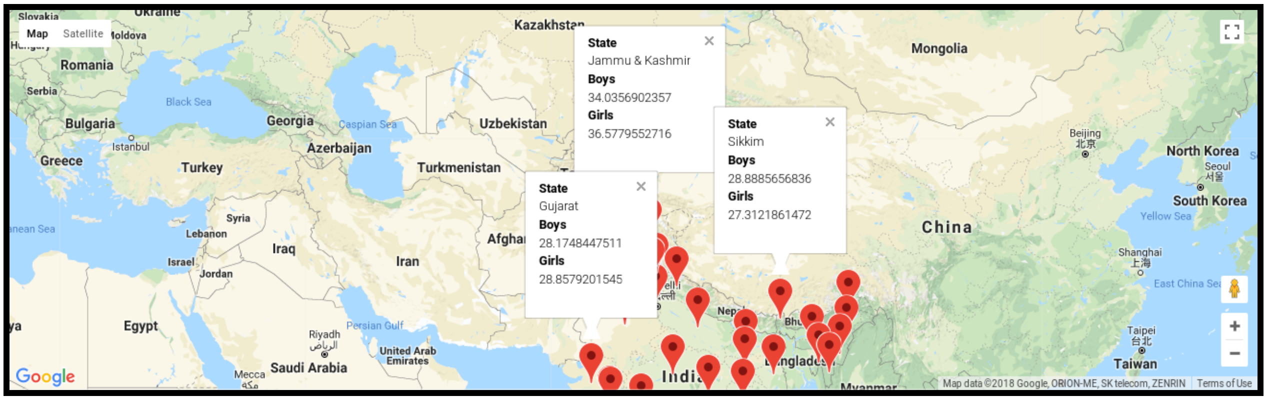
```
Reading %
```

Question 2. How do boys and girls perform across states?

Approach for assessing performance of boys and girls across states

- For these question, only students with gender level either 1 or 2 were selected.
- Then a Pandas PivotTable was created with
  Columns- 'Gender'
  Rows Index- 'State'
  Values- Mean of Overall % (Here mean is considered as here outliers are needed to be considered)
- A google map is created using googlemaps API for creating an interactive map with markers on each state.
- On clicking the marker a window pops up which contains state name, mean of marks of boys and mean of marks of girls.
- Additionally for each state, a bar plot is created, if bar is on left side that shows girls performed better in that state and if bar is on right side that shows boys performed better in that state.

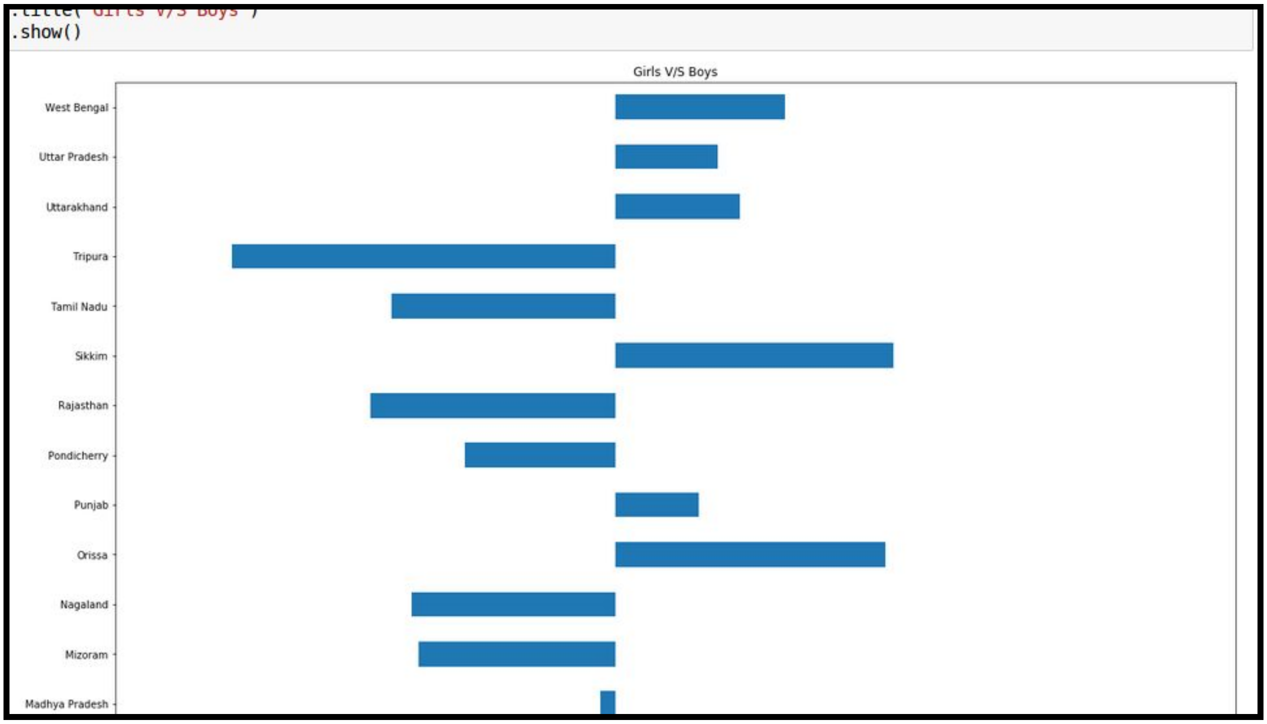
Screenshots-



State
Jammu & Kashmir
Boys
34.0356902357
Girls
36.5779552716

State
Sikkim
Boys
28.8885656836
Girls
27.3121861472

State
Gujarat
Boys
28.1748447511
Girls
28.8579201545



```
.show()
```

Question 3. Do students from South Indian states really excel at Math and Science?

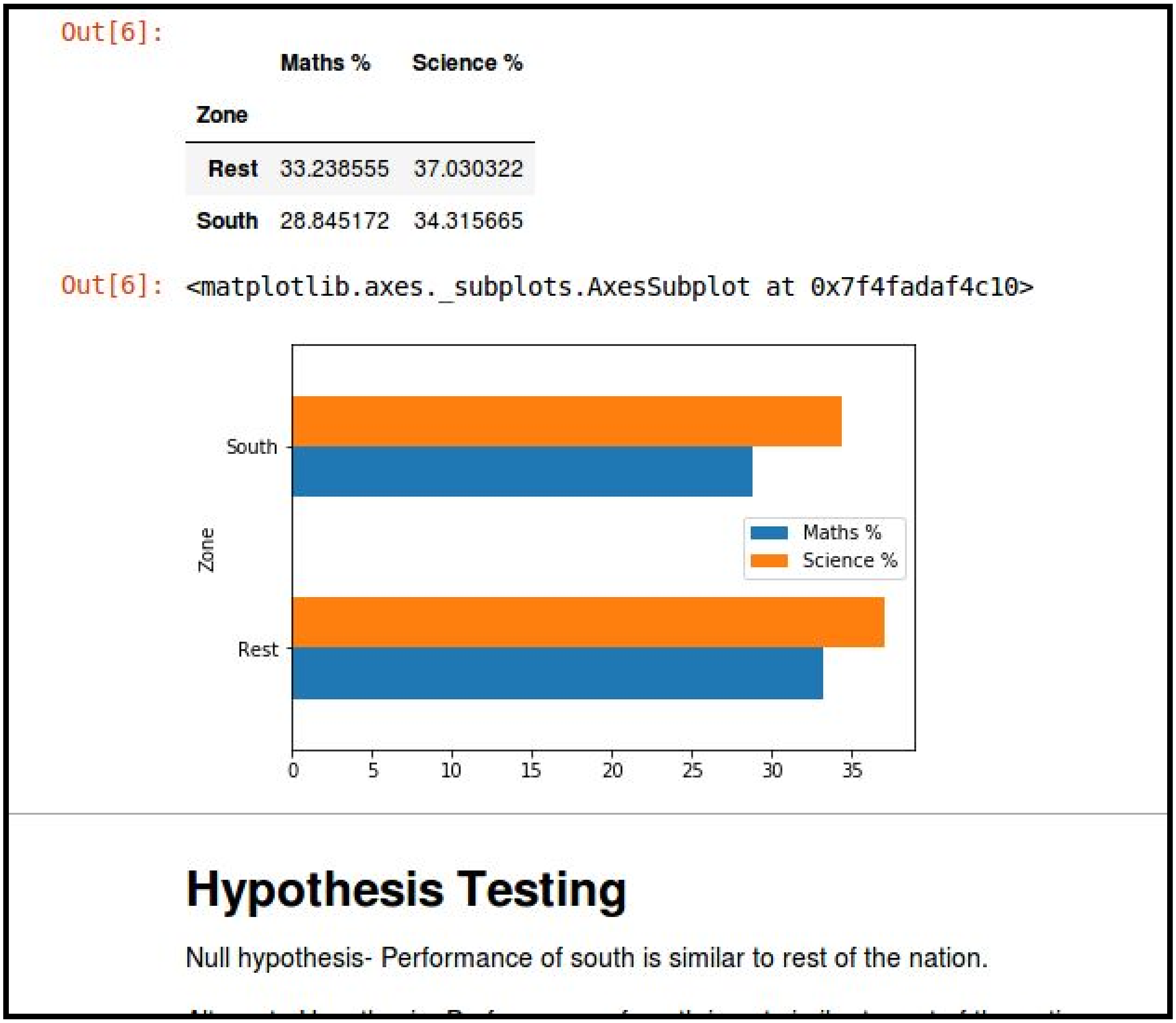
Approach for checking whether students from South Indian states excel at Math and Science

- For this approach data was divided into 2 sets, first set was for students from South Indian states and second set was for students from rest of the nation.
- South India consists of the five southern Indian states of **Andhra Pradesh, Telangana, Karnataka, Kerala** and **Tamil Nadu** as well as the union territories of **Puducherry and Andaman and Nicobar**.
- **HYPOTHESIS TESTING** was performed to check whether students of South Indian states perform better than rest of the nation.
- A **T-Test** was performed with
  Null Hypothesis- Performance of students from south is similar to rest of the students
  Alternate Hypothesis- Performance of students from south is not similar to rest of the students

Significance is set very low as significance level should decrease with increase in number of data points and decrease in variance of the data.

- A **T-Test** makes assumption regarding normality of the data, so firstly it is checked that data on plotting gives a normal distribution curve.
- Once normality of data is assured, a **2 independent sample T-Test** is performed which gives **p-value** and **t- statistic** value as output.
- If significance value is lower than output p-value then NULL Hypothesis can be rejected and based on t- statistic it can be decided whether students from south perform better or other students.

Screenshots-



Out[6]:

Zone	Maths %	Science %
Rest	33.238555	37.030322
South	28.845172	34.315665

```
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4fadaf4c10>
```

Hypothesis Testing

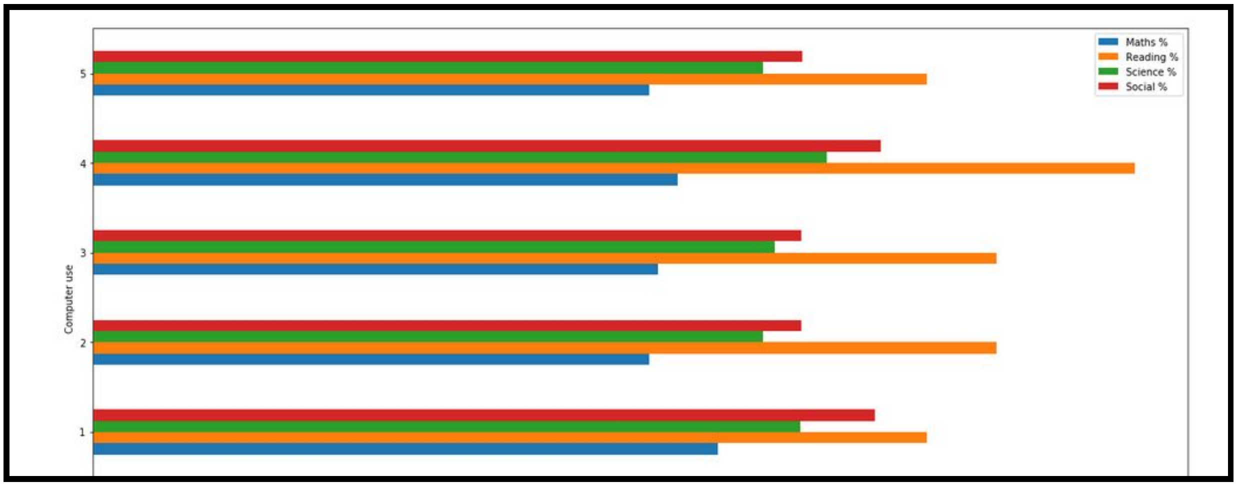
Null hypothesis- Performance of south is similar to rest of the nation.

Question 4. Do students who watch TV and use computer score less than those students who do not?

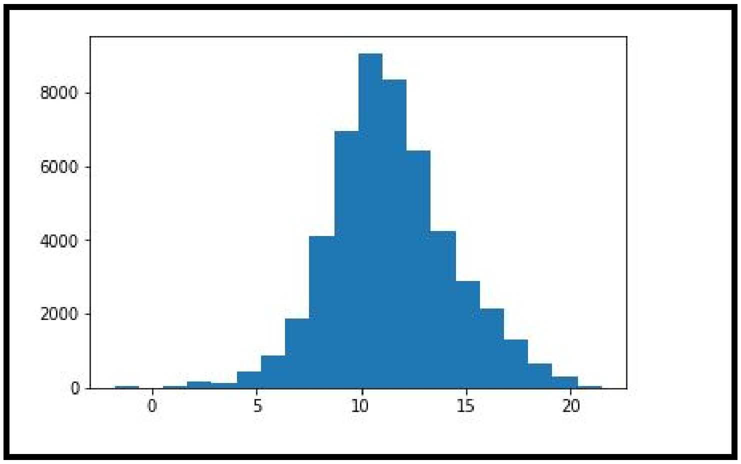
- A bar graph was plotted for mean of marks for students who use computer daily, who use computer once in a week, who use computer once in a month, who have computer but do not use and who never use a computer, this gave some insights regarding effect of using computer.

- Another bar graph was plotted for mean of marks for students who watch TV daily, who watch TV once in a week, who watch TV once in a month and who never watch TV, this gave some insights regarding effect of watching TV.
- Then a **HYPOTHESIS TESTING** was performed with
  Null Hypothesis- Marks of students using computer and Marks of students not using computer are nearly same.
  Alternate Hypothesis- Marks of students using computer and Marks of students not using computer are not nearly same.
  Significance is set very low as significance level should decrease with increase in number of data points and decrease in variance of the data.
- In this case, after dividing data into 2 sets(one for students who use computer and one for students who do not use computer), it was discovered that data does not follow normal distribution.
- In order to convert data into normal distribution **BOXCOX TRANSFORMATION** was applied with different sets of parameter values to **transform data into a normal distribution**.
- The output of boxcox transformation was used for performing **T-Test.**
- **T-Test** was performed using values returned by boxcox transformation function.
- Based on the p-value and t-statistic decision was made whether using computers improves performance or reduces marks.
- T-Test was performed for each subject individually.

Screenshots-



After BoxCox Transformation-



Question 6. Classification model

- Firstly following categories were created on basis of marks for each subject
  High scoring students(HSS) - 85+
  Good scoring students(GSS) - 70-85
  Average scoring students(ASS) - 55-70
  Below average scoring students(BASS) - 40-55
  Low scoring students(LSS) - <40
- One Hot Encoding was performed for the features as they were categorical in nature.
- Then some data cleaning was performed.
- A test train split was created and a random forest classification model with 1500 trees was created.

Submitted By-
Suyash Saxena
suyash2896@gmail.com
+91-9582165990