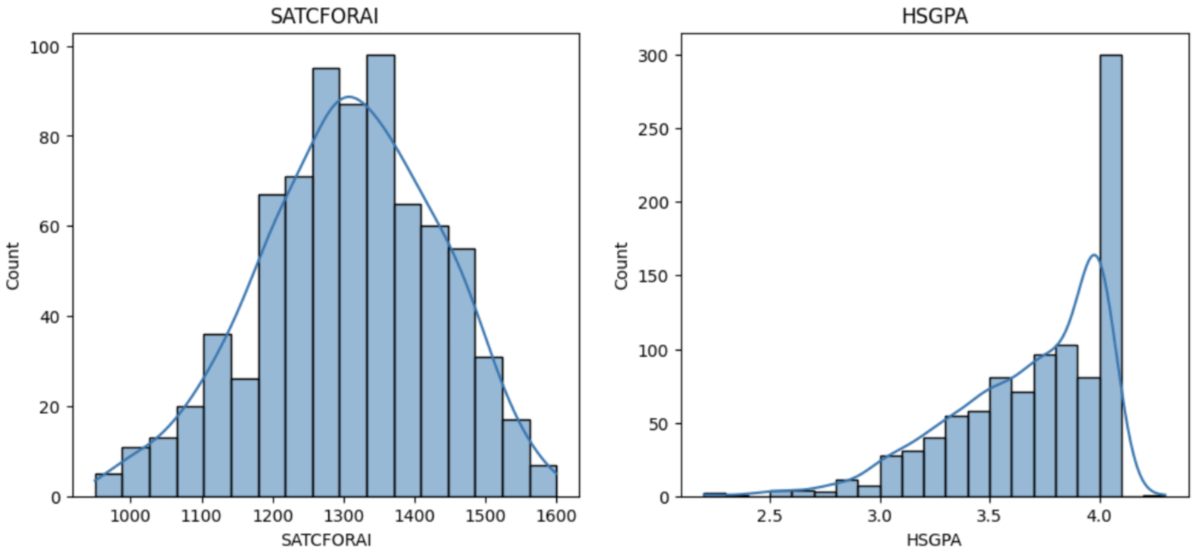
SCRUM REPORT - 1

CSV is cleaned up and converted into a pandas dataframe. We got rid of some unnecessary columns for our purposes i.e. ‘Student’ and ‘US Region’. We also got rid of columns with 0.0 GPA values and we renamed some columns for clarity.

Next, we will work on resolving the ambiguity surrounding “Non-Resident Alien” in our dataset because they can be a non-majority race in their country of origin. We will also start to play around with clustering on different data subsets. We’re not sure exactly what we’re gonna play around with but GPA and seniority might be a good starting point.

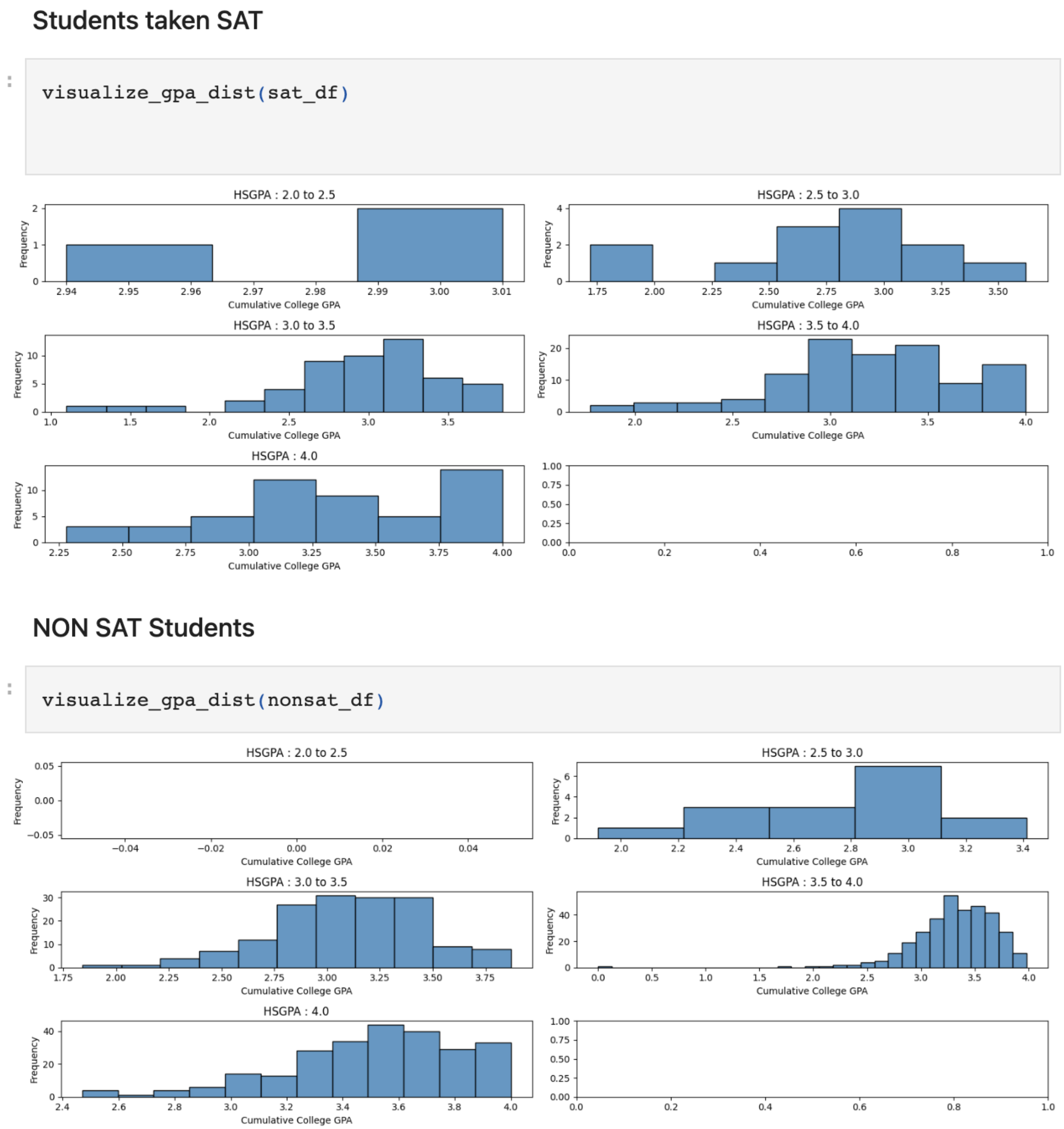
We worked on Data visualization, where we performed distribution of SATCFORAI and High School GPA (HSGPA). The visualization can be seen in the below figure(1).

Count represents the number of people and the horizontal axis represents the SAT scores and the high school GPA’s.



Next we split the data into students who gave their SAT exam and and students who did not give their SAT exam. Later we split the high school GPA into buckets with a bucket size of 0.5. Followed by, we plot the college GPA of the students on the divided buckets.

A clear visualization of the above steps can be seen below:



Students taken SAT

```
visualize_gpa_dist(sat_df)
```

NON SAT Students

```
visualize_gpa_dist(nonsat_df)
```

Explaining the objective of the above steps by explaining one of the graphs.

In the first graph the students with a school GPA from 2.0 to 2.5 have an average college GPA of 2.9 to 3.1. The same is represented for all the graphs with respect to NON-SAT students and SAT TAKEN STUDENTS.