

Names: Carsten Schaufert (C61306769, carstes@clemson.edu)

Lindsey Myers (C76524020, lmyers3@g.clemson.edu)

Link to GitHub: [Project](#)

Background and Motivation

As seniors studying computer science, we are thinking about our futures more than ever and the next steps for us after graduation. A typical post graduation route is finding a full time job, and as we are active in our job searches we wonder “What qualities of a student are good predictors for career outcome?” In this project we explore this question with a focus on career outcomes in the technical industry by using salary as a metric for career outcome. We hope to find meaningful correlations between attributes in our dataset and salary. Our research should help us to better understand how to be successful in our field. We are especially drawn to this project because our field has a massive scope of job opportunities. In order to narrow down what jobs we want to pursue in the future, this project will show us what to expect in our field of jobs.

Project Objectives and Questions

There are many factors that can influence salary outcomes, our mission is to choose a few and observe their relationships to salary. We are especially interested in how different personality traits may affect salary outcomes.

A curiosity we have is about the relationship between gender and salary. We explore their relationship by exploring the salaries of the two genders between various STEM specializations. The aim of this is to observe if any of the genders are favored among the various specializations. Another relationship we wished to explore was if personalities affect salary differently among genders. It would be beneficial to explore these questions in order to observe possible gender biases that may affect salary outcomes.

Personality is a very interesting characteristic for us to look at as it defines who we are to ourselves and the world around us. The broadest problem we look at is how personality affects salary outcomes. We wonder: “Which personality traits and what range of personality scores is most advantageous in determining a successful career?” The benefit of exploring this problem is deciding whether our own personality traits may be putting us at a disadvantage in our job hunts. Additionally, we found it interesting to look at which personality traits may be advantageous by STEM specialization. This is to understand which specializations may be more appropriate for us based on our current personalities. If a person is more introverted, they may want to steer away from specializations that require the use of frequent communication skills.

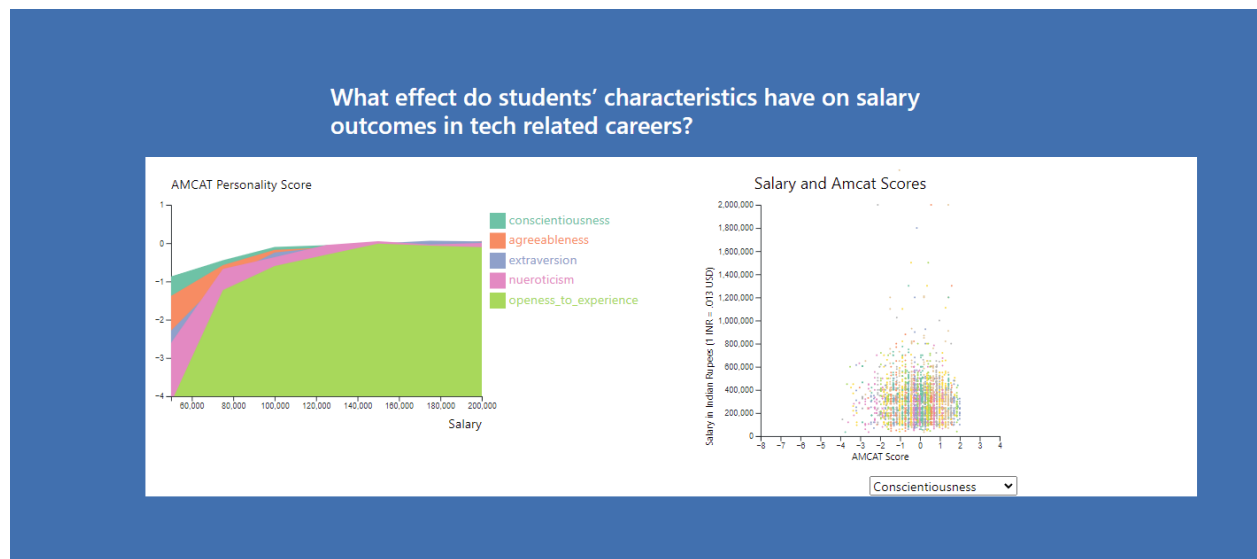
The Dataset: (**Bolded** words indicate a new column)

AMCAT is one of the largest job seeking test services in India, and could be compared to a job seeking version of the SAT, so this score will provide a standardized method of measuring each candidate’s personality traits and their intelligence. Each person in the dataset has a key value **ID** in the ID column. We will be including the IDs associated with the records more so for trackability and to ensure that records are

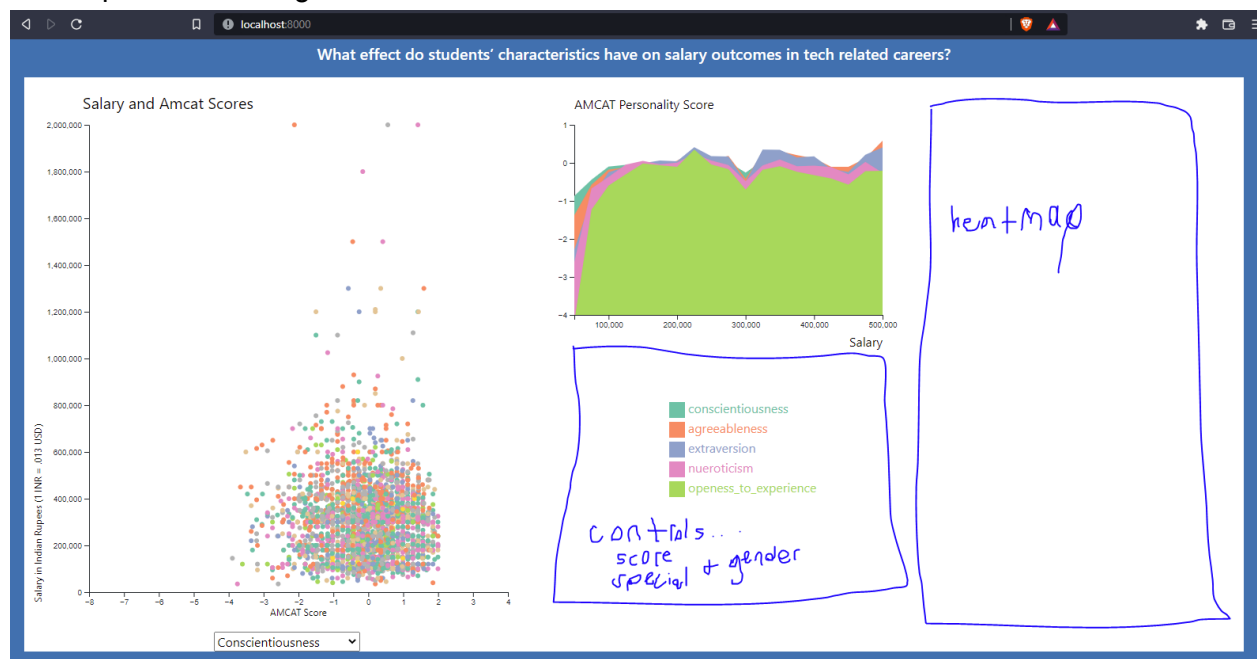
unique instances. The **salary** column is important for our data visualisation. Salary is the dependent variable we are attempting to observe in relation to personality. It can be an indicator for career success, but there are also so many more indicators that are not included in our chosen dataset. For our problem we are only focused on salary as an indicator. We also want to keep an eye on whether **gender** may have a confounding effect on our outcome, as wage gaps exist and may have differing effects across cultures. The **specialization** column includes a person's specialization in their program. Different specializations may have different ideal personalities for candidates and also salary outcomes. For the personality based data columns (Conscientiousness, Agreeableness, Extraversion, Neuroticism, and Openness) a negative value correlates to the person's personality being more like the opposite of the category while a value near 0 means they are in between. **Conscientiousness** represents their willingness to do work and to do it thoroughly, thus if it is negative they tend to be more lazy than conscientious. **Agreeableness** is a measure of how agreeable the person is, so someone with a low or negative score will be more argumentative while a high score represents someone who is agreeable and not argumentative. **Extraversion** is the portrayal of how outgoing and sociable a person is with its negative value associating with introverted more people. **Neuroticism** is an illustration of the mood stability of the person so a higher value represents someone who tends to be irritable, angry, anxious, depressed, or self conscious while a negative shows the person is more iron willed. **Openness** to experience accounts for how likely someone is to try new things and how open minded they tend to be while a negative value represents a more close minded or hesitant person.

The extreme cases where the salaries exceeded 1,000,000 were removed, and the specializations have been generalized because initially there were many which were very similar which had only a small amount of data sets.

Design Evolution

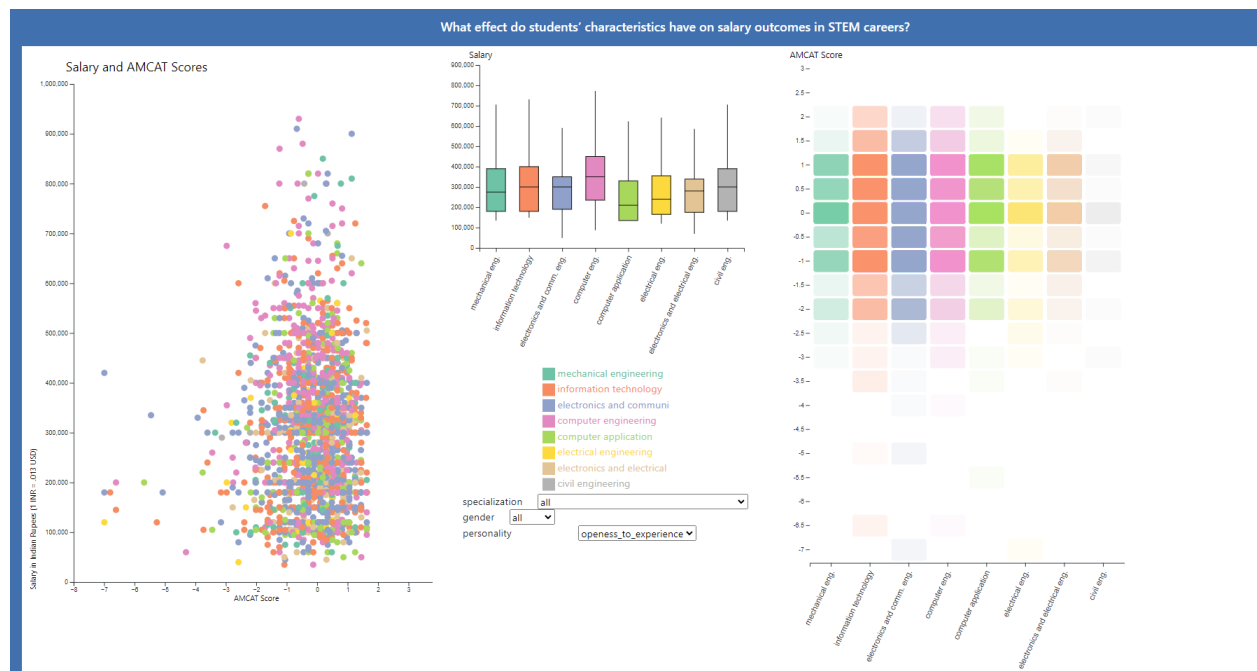


This was the initial iteration. We realized fairly quickly that this visualization was too elementary and did not allow for the user to properly explore the information. The scatter plot was also too small and didn't really say much about what was going on, and the only real interactivity was the ability to change between AMCAT scores. With these issues in mind, we brainstormed and came up with this design.



With that design in mind we changed the middle graph because we felt as though it didn't really convey any information, so we converted it into a box and whisker plot. We also decided to limit the maximum salary to 1,000,000 to allow for the scatter plot to be more readable. We changed the AMCAT score selection to be a drop down to make it so it is easier to select and added a key for the colors and specializations. Furthermore we added the ability to view the data based

on gender and specializations. Finally, we added the heatmap because we felt as though it best conveyed which of the personality traits were most common with each specialization. Below is the final iteration.

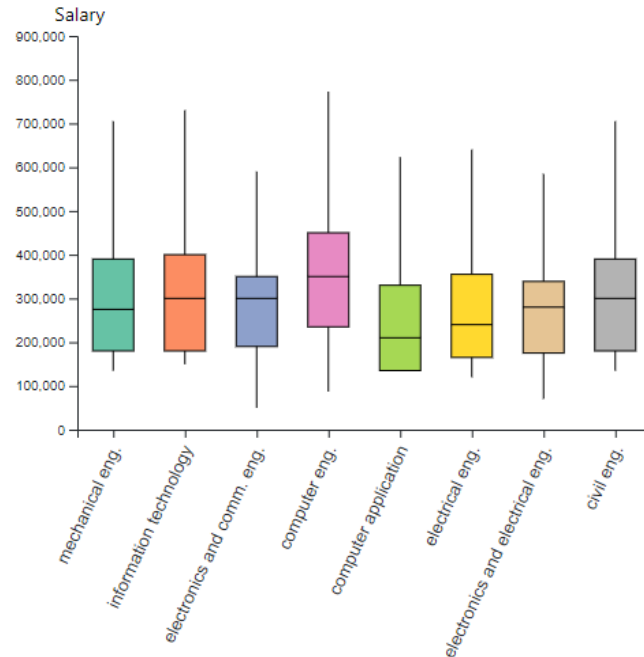


Implementation:

Our main goal for this project was to view the relationship between personality traits and salary outcomes, so it was important to us that the visualization provided the ability for comparison. Our method to enable comparison was to let the user choose a specialization to compare a currently selected specialization to by hovering over the desired specialization in the legend. The colors in all the graphs also correlate to the specialization and are consistent from one graph to the other. The chart will then display the points on the scatter chart and blocks on the heatmap in addition to what specialization is currently selected. This became especially important in comparing salaries among specializations as some specializations had much higher salary outcomes than others.



We decided to include a box and whisker plot to display the ranges and median salaries for the specializations so the user will keep in mind that specializations have different salary outcome statistics. The box and whisker plot is purposefully static on the page so that the user will not forget this relationship while exploring the relationship between personality and salaries.

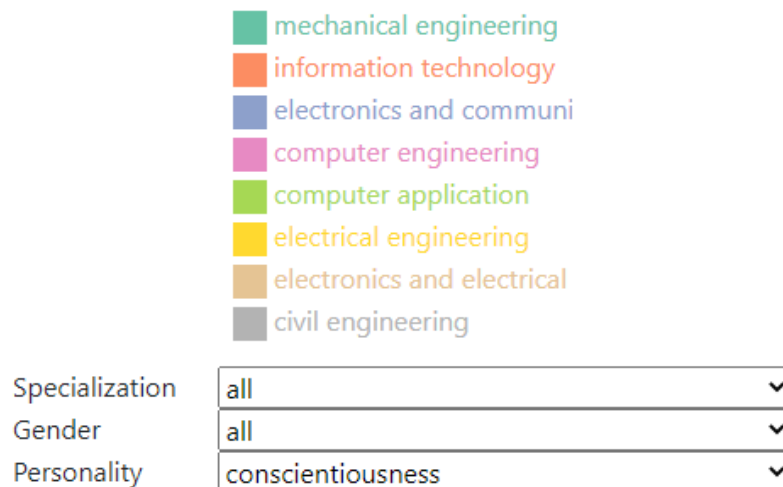


The heat map is useful to convey the relationship between salary and personality using color, while simplifying the view from the scatterplot. As the relationship grows stronger the color saturation does too. The heatmap also updates data in the same way that the scatter plot does.



For interactivity, we provided a few options for the user to select from, specialization, personality traits, and gender. The user can choose which to explore and select from a drop down menu that will update the scatter plot and heatmap; and isolate data the user wants to view. The user can also display a specialization to compare with

the currently selected one by hovering over a specialization in the legend. Once the user selects a specialization the scatter plot's colors change to correlate to the gender allowing the user to quickly analyze the data.



Specialization	all	▼
Gender	all	▼
Personality	conscientiousness	▼

Evaluation:

Through our visualization we were able to observe that conscientiousness, agreeableness, openness to new experiences, and neuroticism are all characteristics that are favorable to have more of for higher salary outcomes. This trend is pretty common among all of the specializations and is displayed through the heat map when all specializations are selected. Individually looking at specializations on the scatterplot, you can notice that those with lower scores for any personality trait almost always had lower salary outcomes, but those with higher scores may or may not have higher salary outcomes. The points in the scatter plot make almost a clear triangle in each of personality traits which helps make it clear what the correlation is. We also were able to view the drastic differences between the amount of men compared to women in STEM fields. An issue was caused by this, as the female's data at times became too sparse. We were also curious to see what specializations were likely to have higher salary outcomes and this was very perceivable from the box and whisker plot and when breaking down the scatter plot by specialization.

We feel that the visualization answered our questions but it could use some improvements. An issue we found is that the legend interaction is not very intuitive and the user has to discover it. A way to improve this may be by giving the user a hint to hover over the legend and to allow the interactivity when the user hovers over the label. It would also be a good idea to allow for the box and whisker plot to be filtered by personality traits also as this is interesting data to have. Another issue we have is when points on the scatter plot have the same value you cannot tell that there are multiple points there.