<u>Academic Impressions Applicant Analysis</u>

Introduction:

We located a company in South Denver that specializes in creating training software curriculum for professionals in academia. They are a medium-sized technology company and currently experiencing rapid growth. Their Human Resources Manager provided us with a dataset of applicant tracking data from the last ~15 months. Our stakeholder originally gave us this dataset because she wanted to see what interesting trends and patterns we could find. She had one specific question of the information which we will add, but overall wanted to see what we would come up with in our analysis.

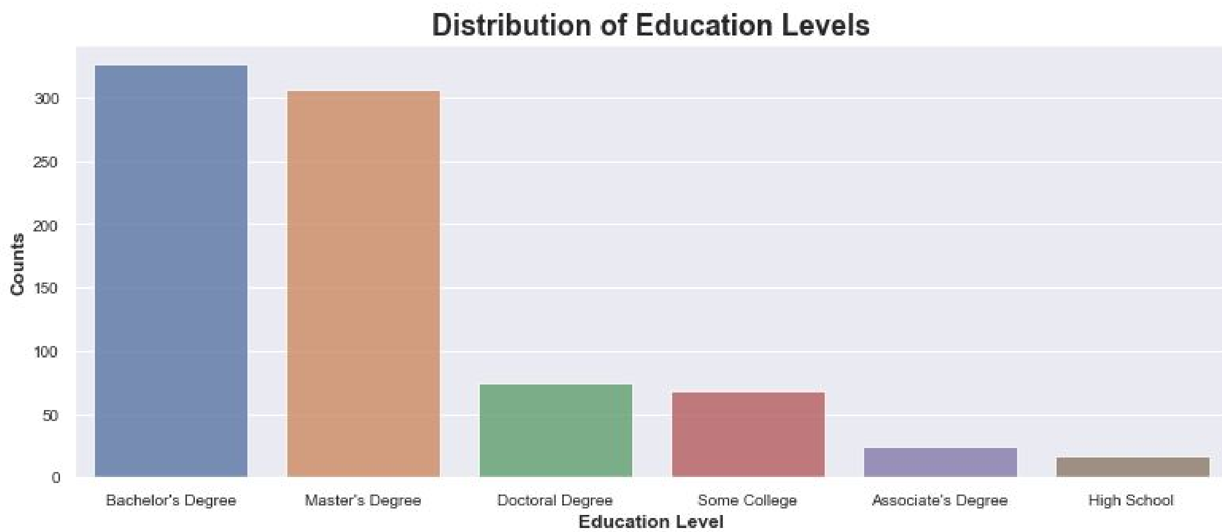
> ***Stakeholder Question:*** *Why are our doctoral candidates are not getting offers or dropping out during the interview process*?

The original dataset we collected from our stakeholder had information on the applicants of the company. Features included: application date, first name, last name, position, department, location, employment type, current status, education level, desired salary, and source. Our first step in cleaning the data was to drop the unneeded columns: *department, location, employment type, and source*. Next, we decided to regroup different subcategories within features to make the data simpler and easier to interpret. For example, in the 'Education Level' column there were several different variables for a bachelor's degree, we ended up combining all the variants into one variable. To answer questions concerning the application date, we had to convert the variables to a date time value using a built in pandas feature that is offered within our programming language - Python. Later in our analysis, we introduce two new datasets. One dataset collected information concerning salaries at AI, and the other data concerning national salary averages across the country via Glassdoor.com. Then we fused these two new datasets to compare salary ranges. We asked several questions of the data, and answered them in various ways using Python and visualizations offered within Python.

Analysis

Question 1: What does the distribution of education levels applying to Academic Impressions look like?

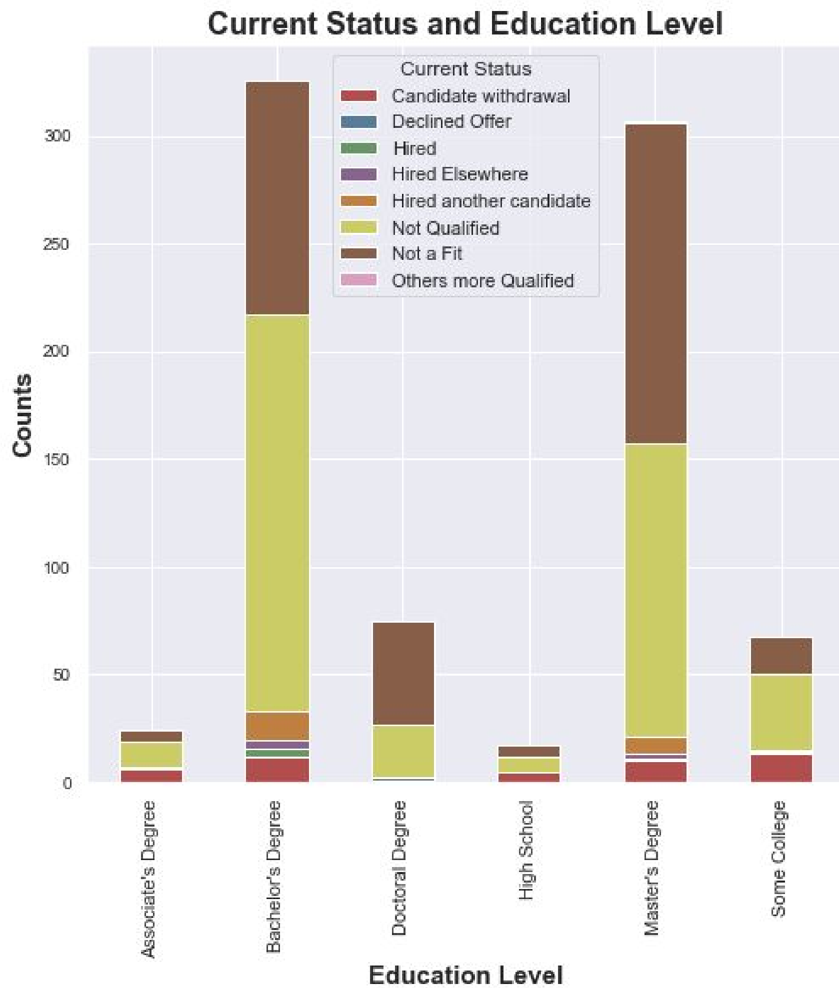
To answer this question we first needed to gather a counts data frame concerning the education levels. Once completed, we used it to create a bar plot and visualize the distribution.



From this graph, it may look like the reason Doctoral candidates are not getting hired is because of the sheer number of Bachelor/Master candidates. We were told that Doctoral candidates are preferred and that education is something that is highly regarded at Academic Impressions.

Question 2: What is the spread of current status according to education level?

To accomplish this, we had to create a crosstab of our original dataframe grouped by two columns we are interested in. From this, we created a stacked bar chart visualizing our distribution.



This yields a distribution that you would expect from the current status of these education levels. One thing to note is how more than half of the Doctoral candidates are categorized as 'not a fit'; something that is not true of other education levels - more on this later in our analysis.

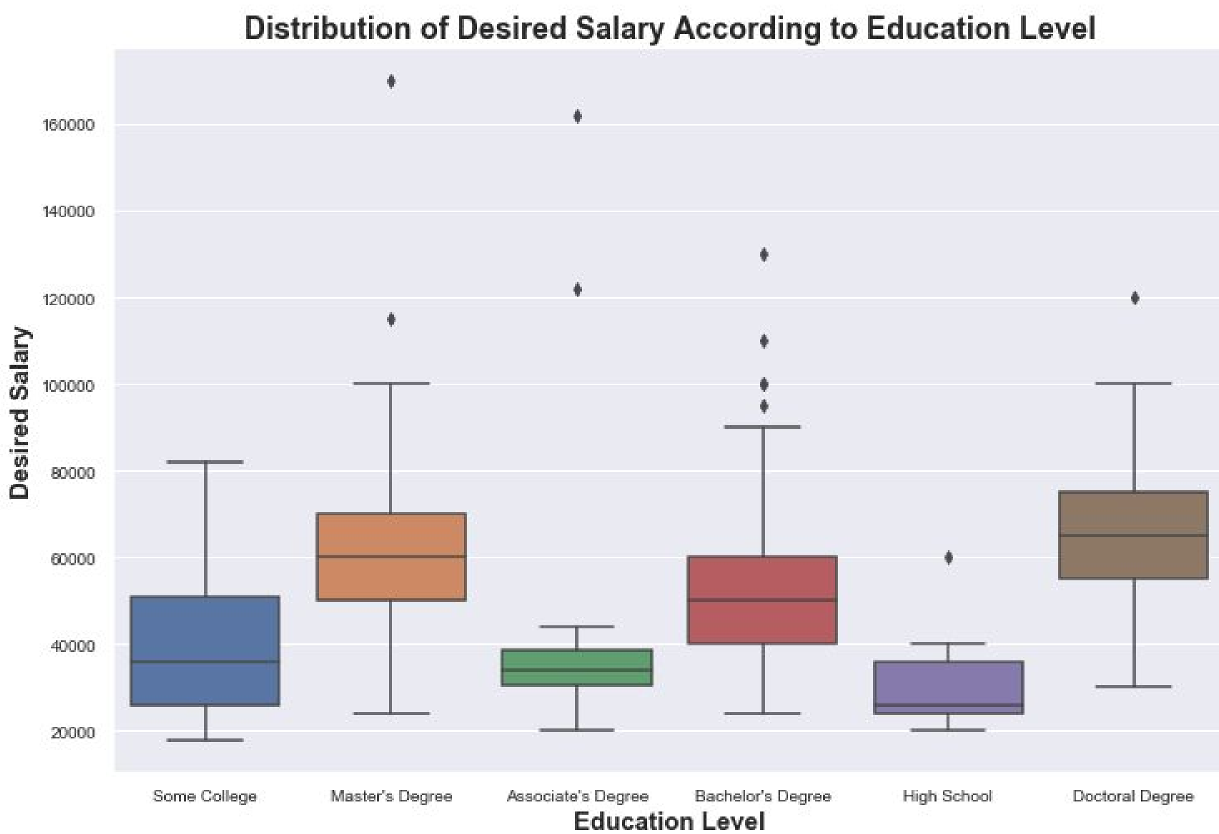
[Main Objective] Question 3: Why are doctoral candidates not getting hired, what does the data say about this?

We based our entire study on this question and most of our analysis as this was one of the specific questions that our stakeholder had of the provided dataset. We attempted to analyze the data on these candidates to see why they were slipping through the cracks in the interview process. To start the process we isolated just the Doctoral candidates into a separate data frame. Using this new dataframe we found the distribution of positions being applied for.

```
doctorate_df["Position"].value_counts()
```

```
Research Analyst           59
Instructional Designer     14
Journalist                  1
Associate Brand Manager     1
Name: Position, dtype: int64
```

From this we see that the majority of doctoral candidates are applying to be research analysts. Next, we attempt to visualize the distribution of desired salary to compare what doctoral candidates are desiring, to everyone else.



Looking at this boxplot above we can see that the doctoral candidates are desiring a higher average salary than other education levels. This could be factoring in to why they are not getting offers.

4 Findings of Doctoral Candidates:

- A lot less of them are applying compared to other education levels.
- More than half are categorized as ‘Not a Fit.’
- They are desiring higher salaries.
- Applying to be primarily Research Analysts.

In a pure numbers game, they are being overwhelmed by Bachelor's and Master's candidates, even though they are considered top tier candidates. Doctoral candidates are not fitting within the organization (which will hopefully make more sense to the HR Manager). Their higher salary range makes it tough to hire them over an equally qualified candidate. Lastly, they are largely applying to be Research Analysts; we are not certain what to do with that information, but combined with the other findings it could be an important trend for our stakeholder to further examine.

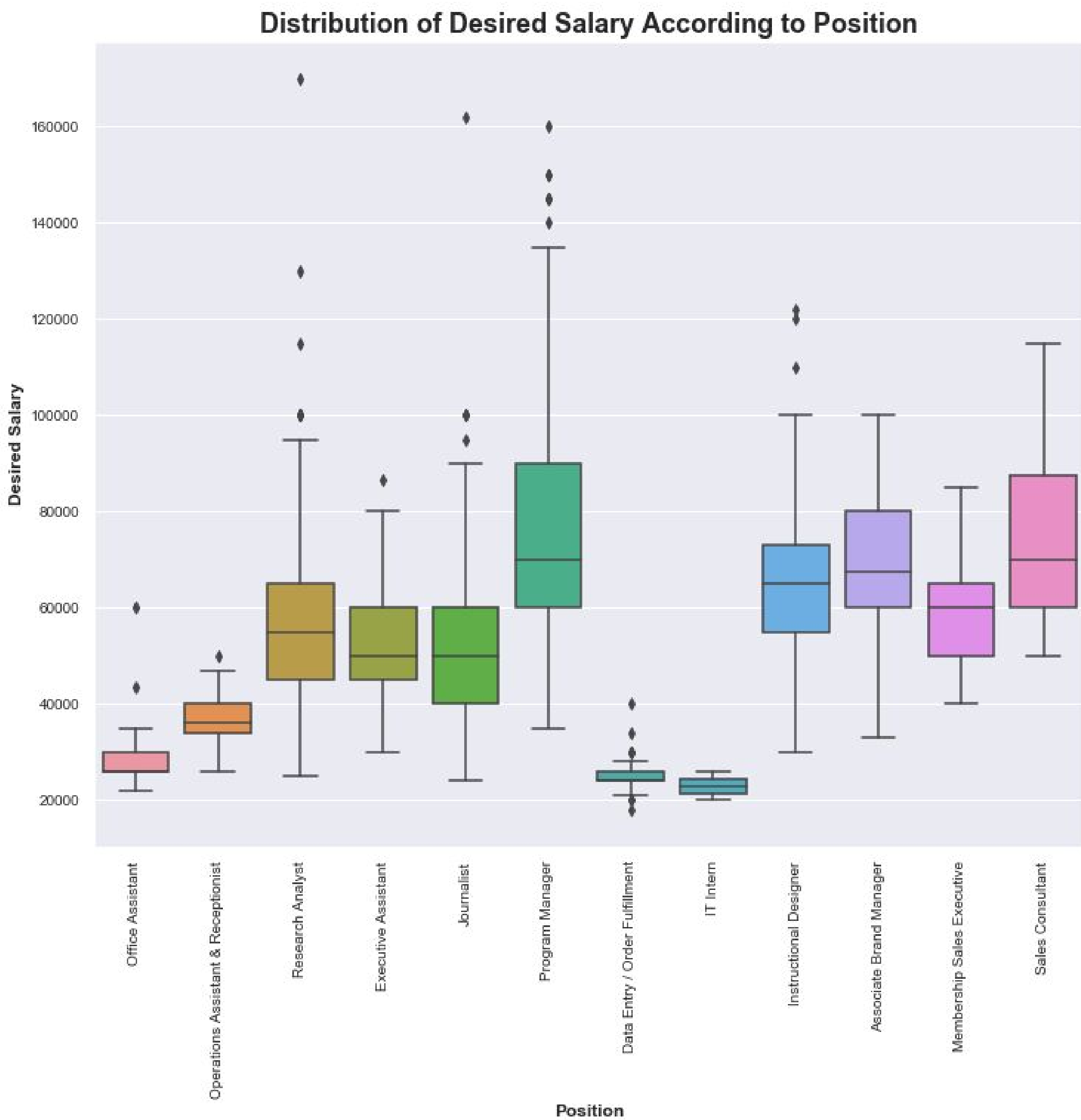
Question 4: What is the most popular position according to applicant numbers, and which one has the highest desired salary?

To start this one off, we find the value_counts() of the position levels in our dataset.

```
df['Position'].value_counts()
```

```
Research Analyst                      428
Program Manager                       161
Instructional Designer                153
Journalist                             97
Executive Assistant                    78
Data Entry / Order Fulfillment         58
Operations Assistant & Receptionist    52
Office Assistant                       51
Associate Brand Manager                34
Membership Sales Executive             17
Sales Consultant                       15
IT Intern                               2
Name: Position, dtype: int64
```

The dataset we were given is overwhelmingly research analyst applications. This bias in the data will skew our results slightly when doing further inferential and descriptive analysis, and our stakeholder must keep that in mind. It's also an interesting piece of information that our stakeholder might not have been aware of as to why are so many candidates are applying for this role? After this, we mapped out the distribution of desired salary according to position using a boxplot.



From this boxplot we can see that the top three positions in terms of desired salary are: Sales Consultant, Program Manager, and Associate Brand Manager. Are these positions' desired salary ranges reflective of their responsibilities and workload in the company?

Implementation of Second and Third Datasets (Sections 5 and 6)

- To compare the data we have from AI with outside information, we will introduce a second dataset from Glassdoor.com concerning positions and their respective salaries.
- Upon acquiring the dataset from Glassdoor, Academic Impressions provided six salaries of positions in their company that were in the Glassdoor dataset. We fused this information with the data from Academic Impressions to create a third dataset.

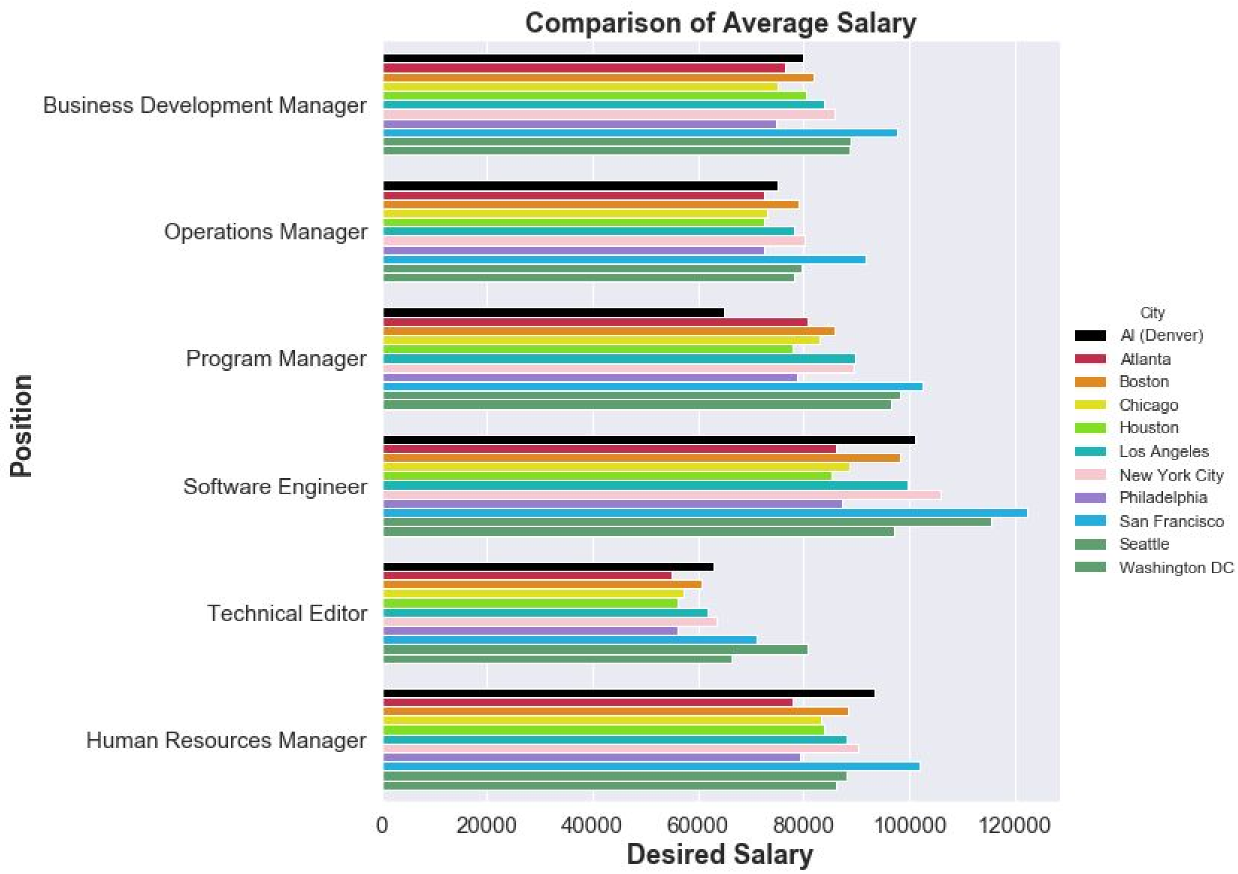
Question 5: What are the national salary averages of job titles that Academic Impressions currently employs?

After cleaning our newly created dataset, we ended up with 3 features: city, job title and salary. At this point we felt that creating a national dataframe combining salary information from other cities around the US might show if AI is paying below market value for the positions they frequently hire for. We were able to match six positions between Glassdoor and AI and were given the starting hiring salary of those positions.

	Job_Title	Salary
0	Business Development Manager	83057.272727
1	Human Resources Manager	87320.727273
2	Operations Manager	77521.000000
3	Program Manager	86163.454545
4	Software Engineer	98856.636364
5	Technical Editor	62864.000000

Question 6: Is Academic Impressions paying above or below average compared to to other cities represented in our new dataset?

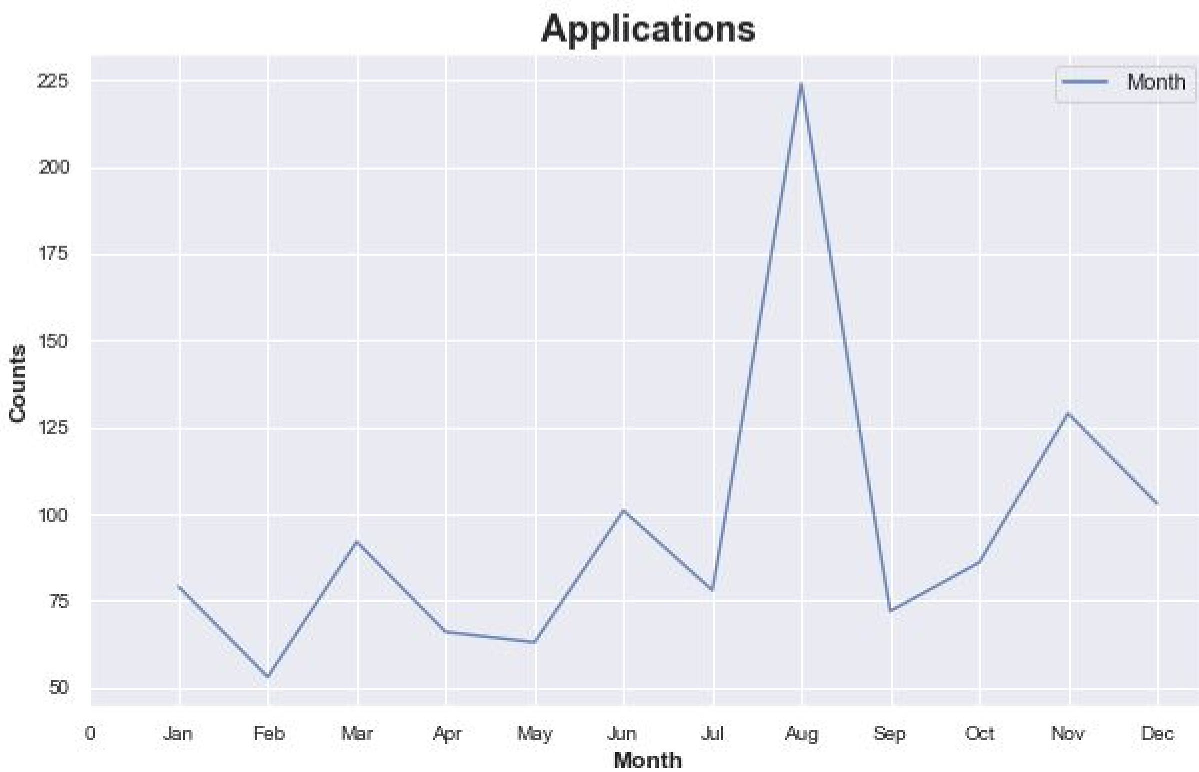
We then created a horizontal bar chart visualizing all of this newly acquired data.



From this we see that AI s is paying around average, if not better in some categories, compared to other cities.

Question 7: When are there spikes in applications?

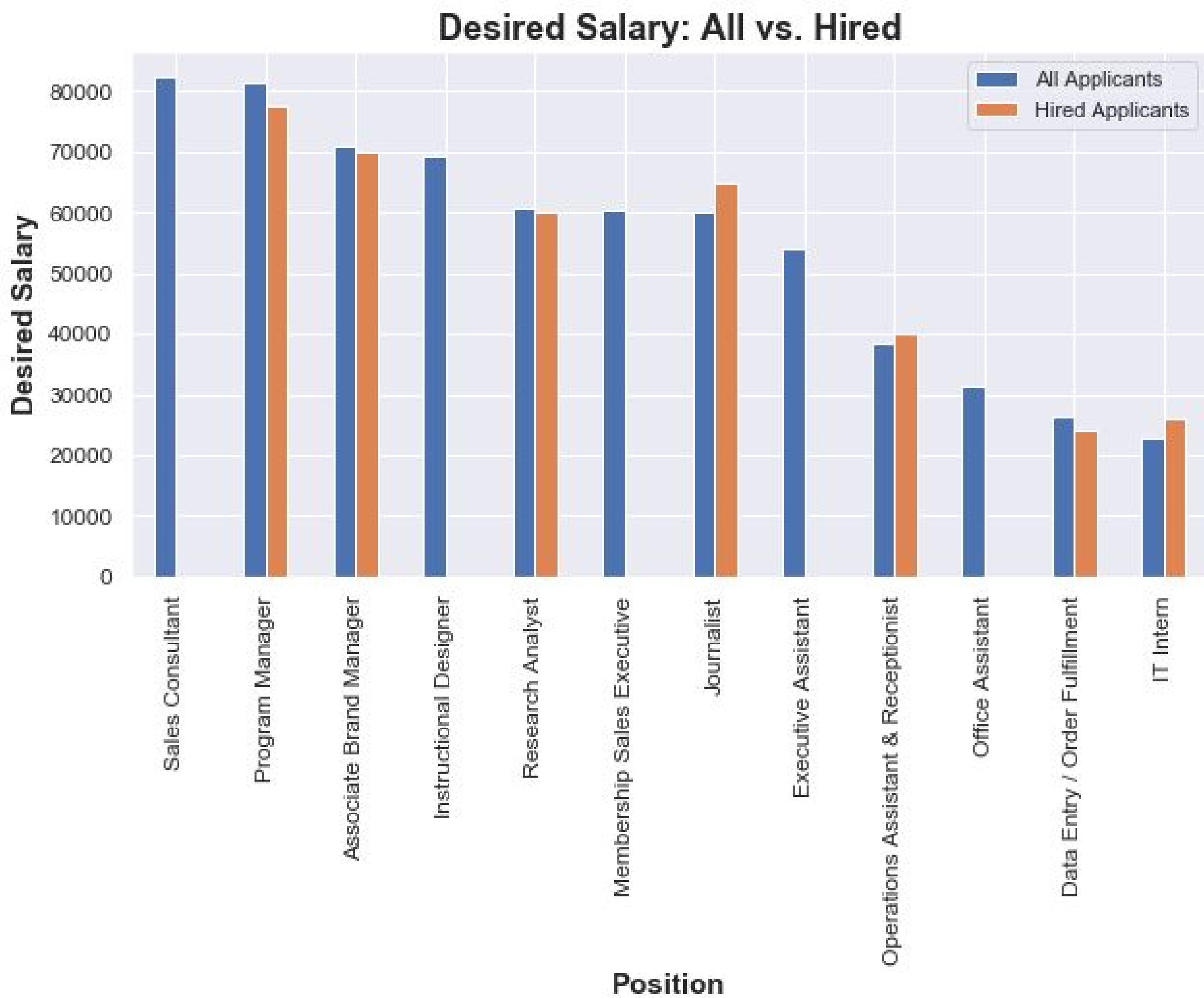
In order to answer this question effectively, we looked at the number of applications per month according to our dataset and mapped it to a line graph.



As we can see from the graph above, there is a major spike in applications in the month of August. Not exactly sure what is causing this. This is useful for an HR manager to know so they can better prepare for a greater influx of applications if this trend were to continue.

Question 9: What sort of candidates are getting hired?

First, we isolated only the candidates that were hired into a seperate dataframe. After looking through these 12 hired candidates, it was interesting to see that only 1 person was hired of the 428 that applied to be research analysts. We attempted to visualize the desired salary of everyone in the dataset compared to the hired applicants to see if their numbers were lower.



This chart shows us that they were right on par with the rest of the group. Overall, we weren't able to effectively find out what sort of candidates are being hired given our dataset.

Question 9: Are we able to predict position based on education level and desired salary?

To find this out, we attempted to initialize a logistic regression with education level and desired salary as features and position as the label. After vectorizing the education level feature and fitting the model, we ended with 51% testing accuracy. We initially thought this was surprisingly good given the little tuning we did for the model. After diving deeper into the predictions, we found out that our model was guessing the position to be research analyst every time. Because most of our applicants were applying for this position our model decided to predict the label every time and get the best accuracy possible. Ultimately, this logistic regression could be interesting to our stakeholder, but our initial model is not fine tuned to fit the data well enough yet. If fixed up sufficiently it could be very interesting to have high accuracy in predicting position based on desired salary and education level.

Conclusion:

We believe that we could not accurately answer Academic Impressions' original question of "Why Doctoral candidates were not getting hired or dropping out of the interview process?" We have sufficient evidence as to what is happening with the applicants in the interview process, but we lack the details that make up the "not a fit column" in AI's data. If we had access to the hiring manager's notes on each candidate as to what ultimately makes a candidate not a fit, we could have coded this data into categories that best described the hiring manager's notes and used this to help further our analysis.

However, we are able to provide Academic Impressions with other interesting findings that we feel will are important such as the large spike of applicants during the month of August, or how AI's salaries stack up against other cities around the US. In terms of bias, the research analyst position was our main culprit. This position overwhelms our dataset and skews our data as a result. It is also very broad and encompasses a variety of skill sets. An analyst role like this one is going to differ greatly from company to company. It would be doable, but difficult, to make an apples to apples comparison across industries, which might affect salary and give us more insight to answer the original questions from our Stakeholder. In terms of next steps, if Academic Impressions would like to have more insight into the hiring analytics, we suggest a more powerful applicant tracking system that is able to figure out a lot of these types of analytics for you, or AI is going to have to do some of this analysis in-house where confidential information can be more easily shared. We have shared as an attachment to this email our Jupyter Notebook that includes all the code we used along with annotations. This will be a great starting point for any further hiring analytics AI chooses to perform. We are hoping future analysts can build off of what we have found out through our given resources and expound upon it.

Kelly - *Thank you for sharing your information with us and being so quick to respond to emails and phone calls!*

- *Aaron Roberts and Nathan Duffy*