SSC442: Exam 1

Riley Latham

Workforce Analytics:

1. Hiring is an incredibly costly process and lowering turn-over rates can help drastically reduce these costs and keep a strong employee base. Using the dataset to predict who stays with the company longer than some cut-off point at which hiring another or sticking with the current employee nets to a zero cost can help to achieve this. For the sake of simplicity let's say that once an employee has stayed longer than one year with the company they have become "worth" the investment. We can then split our data into those who stay with the company for one or more years and those who are hired and leave earlier than one year. Having split our data we can take a look at their job application similarity, base salary and any bonuses given throughout the year, manager and teams they're assigned to, as well as any moves they made from one division of the company to another. Using these combined data points we can hope to find a relationship between those who choose to stay with the company and those who do not.
2. The results from question one are fairly easy to interpret. The objective is to use job application similarity, salary and bonus data, and manager/team assignment to determine how long a hired employee stays with the company. We have segmented our data into those who leave prematurely and those who stay with the company for a decent amount of time. Using our inputs we're able to find if certain job applications tend to yield employees that stick around or not, how base pay and bonuses affect an employee's willingness to stay with the company, or even find managers and teams that may need to

be targeted for inclusivity training. The results can be acted on to help retain employees and lower the very high cost of hiring new employees.

Echo Team:

1. Seeing as people rarely report an issue using Alexa we won't have good data as to what problems we should be focusing on with respect to voice recognition (maybe Alexa has a hard time understanding words starting with a 'th'). In order to prioritize our coding efforts a visualization might be able to find these same issues without the user needing to report an issue. If an Alexa is asked a question it will ping on, record the sound, and act on this if it recognizes the voice command. We can then identify a subsection of our data where the Alexa is pinged on three or more times in a minute, this should be enough to identify that we have a problem understanding the command being given. Using this subsection of data we can find what problems we're having with voice recognition and work to fix this. The plotting could be done using a histogram from ggplot. First we subset our data to the voice lines that are recorded once Alexa pings on at least three times in a minute. These voice lines should have a similar pattern and we can group all such data. Using these grouped sets of data we use a simple histogram from ggplot to track which commands that are not understood are being given the most often. Then we can target these sounds for more work on being able to recognize them.

Twitch Data Science:

1. It seems like the most useful Y-variable should be the amount of bits donated per hour streamed. This helps to make sure that we don’t bias our model towards people that stream a lot and so get more time to earn bits compared with somebody who streams much less but earns many more per hour. I don’t understand many of the intricacies to streaming but some X variables that would be helpful to track could include the genre of game as well as the actual game they play, perhaps gender, race and nationality play a role, time the stream occurs, and what language they speak. If we have data on their viewers as well using these same demographic break ups can help identify who we advertise the bits to. All of these characteristics would be able to help estimate what streamers earn in bits per hour.
2. You should use around 10% of the current users as an experimental group. This gives enough people to show trend changes from the control group without serious repercussions should the experiment have unintended consequences. It should be randomly split amongst our target demographics described in part one to ensure that we are using randomized data and aren’t producing biased estimates from our experiment.