At the very beginning, place one big table listing each variable that your roadmap says should be collected, and indicate the values and granularity you would want for the variable. Be as specific as you can.

General suggestion: think about whether each of your action questions contributes to the goals in your 3rd progr report slides:

* Validate the hypothesis that competition for qualified candidates has been increasing,including from eCommerce companies

* Propose a roadmap to attract more job applicants and improve retention

If the match is good, great! If the match is not good, how can you modify the action question to make a better match?

Action Questions

Also, of course, indicate whether that variable is already collected adequately by one or more companies.

Maybe also indicate which action question (or questions) that variable would be relevant for.

Time to Fill Requisition

1. **What is the predicted time to fill for job openings in different seasons, positions, and locations?**

Subquestion: How is it likely to fill a position in different seasons, positions, and locations?

Importance: With an estimate of time to fill a position, the company can plan for recruitment more accurately and understand the amount of time to obtain a qualified candidate differs for different seasons, positions, and locations. The company can also establish a benchmark for this metric and track any noticeable changes to detect problems early on.

Variables tracked: Requisition open date, requisition status (closed and filled / closed and not filled), requisition job title, requisition location

Variables not tracked: Requisition close date, requisition wage budget, requisition sourcing channels

Methods of analysis

General comment: in final report/presentation, give examples of using the methods for each question to help answer (or get closer to an answer for) each action question.

i. Exploratory data analysis
   1. Calculate time to fill for all closed requisitions in at least a one-year period
   2. Use mosaic plots to compare requisitions that were filled and not filled for different positions and locations
   3. Use line plots to visualize the number of filled requisitions and the number of unfilled requisitions over requisition open time
   4. Filter out requisitions that were not filled, use boxplots to visualize time to fill for each job position and location
   5. Filter out requisitions that were not filled, use line plots to visualize average time to fill over requisition open time.
   6. Filter out requisitions that were not filled, use boxplots to visualize time to fill for each month.
   7. Segment data by job position, department, location, and repeat the process to examine possible effect of seasonality for different job positions/departments/locations

8. Filter out requisitions that were not filled, use scatter plot to visualize the relationship between requisition time to fill and requisition budget wage

ii. Linear regression model, tree-based models (decision tree, random forest, XGBoost)

1. Use requisition open time, requisition position, requisition location, requisition wage budget, and requisition sourcing channels to predict time to fill a requisition

2. What types of applicants tend to apply early/late?

Add here the idea that candidates who apply late may be more like those that don't apply at all.

Importance: Although the company does not have access to potential candidates who did not apply, we can investigate whether applicants who apply early have different characteristics compared to applicants who apply late.

Variables tracked: Requisition open date, requisition job title, requisition location, applicant application date

Variables not tracked: Requisition close date, applicant sourcing channel, applicant demographics, applicant work history (#years of experience), connection between requisition and applicant data

Methods of analysis

General comment: in final report/presentation, give examples o using the methods for each question to help answer (or get closer to an answer for) each action question.

i. Exploratory data analysis

1. Use applicant application date - requisition open date to obtain the response variable "application days"
2. Use boxplots and line plots to compare the distribution of "application days" for different applicant sourcing channels and for categorical demographic variables such as gender, race, education
3. Analyze the plots and understand whether the differences result from different posting dates in sourcing channels / different time for information in each sourcing channel to reach potential applicants
4. Use scatter plot to explore the relationship between "application days" and applicant age

5. Segment data by job positions, department, location, and repeat the process to examine whether early applicants and late applicants behave differently for these groupings.

ii. Linear regression model, tree-based models (decision tree, random forest, XGBoost)

1. Use requisition job title, requisition location, applicant application date, applicant sourcing channel, applicant demographics, and applicant work history (#years of experience) to predict "application days".

3. Do shorter interview cycles increase likelihood of hires, among those who were offered a position?

in our meeting you also hypothesized that higher quality candidates prefer (or will be snapped up by) companies with shorter interview cycles. point this out here.

Importance: We think shorter interview cycles save costs in revenue by reducing the time to search for quality hires, while also increasing their chances of getting a qualified candidate in this market where the demand is high.

Variables Tracked: Talent Acquisition Interview Bin (one client)

Variables Not Tracked: Interview Bin of Candidate, Interview Date, Status of Interview

Methods of analysis

General comment: in final report/presentation, give examples of using the methods for each question to help answer (or get closer to an answer for) each action question.

i. Exploratory Data Analysis

1. With each column representing the different stages of the interview process, including hiring, obtain the dates of each stage for a given requisition.
2. Calculate the difference between the offer date and the application date for a candidate, for a given requisition. Keep count of whether the candidate accepted the offer or not.
3. Aggregate the differences between the offer date and application date (in days) for candidates who accepted the offer.
4. Aggregate the differences between the offer date and application date (in days) for candidates who did not accept the offer.
5. Split the length of the interview cycles (the difference between application date and offer date) into the number of bins you see fit: 1 month, 2 months, 3 months, …, etc. These bins indicate the duration of the interview process in order for us to observe the offer acceptance rate.

6. For each bin, calculate the offer acceptance rate and plot a bar graph displaying the offer acceptance rate for each bin, which indicates how long the interview process was for them. You can also observe the most common time it takes to complete the interview cycle and the average time for the interview cycle.

I like these category labels. Maybe make them just a little more prominent - I missed them on my first time through the document.

Sourcing

4. **Which sourcing channel provides the candidates that are most likely to get hired?**

great

   a. Importance: With this information, one can choose to spend more effort on the most efficient source of hiring to speed up the recruiting process. Assumption: Hiring status happen within the same period as requisition opening/application Although we have date of hiring, it is more reasonable to use requisition opening/application date to link back as "for requisitions opening at certain period, what are their hiring rate". Assumption: Each position is subject to equal opportunity of source of hiring.
   b. Variables Tracked: Application Date, Division, Requisition Open Date, Status
   c. Variables Not Tracked: Source of Hiring
   d. Method of Analysis: For each source of hiring, calculate the hire rate for each period of time.

General comment: in final report/presentation, give examples of using the methods for each question to help answer (or get closer to an answer for) each action question.

5. **Which sourcing channel provides the candidates that are most likely to stay for more than a year once hired?**

great

   Importance: Answering this question helps the company understand which sourcing channel provides candidates with the best quality. The company can put more resources into effective sourcing channels and perhaps drop some lagging recruiting strategies.

   Variables tracked: applicant disposition, employee tenure

   Variables not tracked: applicant-level sourcing channel, employee performance metric in the first year, connection between applicant data and employee data

   Methods of analysis

General comment: in final report/presentation, give examples of using the methods for each question to help answer (or get closer to an answer for) each action question.

   i. Exploratory data analysis
      1. Calculate source of hire (the percentage of hires entering the pipeline from each sourcing channel)

2. Use mosaic plots to compare the percentage of dispositions for each sourcing channel
3. Analyze the plots and understand whether the differences reflect biases in the recruitment process
4. Filter out rejected/withdrawn candidates, use boxplots to visualize the performance metric for each sourcing channel
5. Filter out rejected/withdrawn candidates, use boxplots to visualize tenure for each sourcing channel (Be mindful of newly employed sourcing channels)
6. Segment data by job position, department, location, and repeat the process to examine whether the effectiveness of sourcing channels differ for job positions/departments/locations
7. Consider grouping sourcing channels or ungrouping sourcing channels if differences/similarities are observed

ii. Linear regression model, tree-based models (decision tree, random forest, XGBoost)
   1. Use only sourcing channels to predict employee tenure and employee performance

iii. One-way ANOVA/ANCOVA / two-way ANOVA (sources, hired or not)

Termination prediction (post-hire)

6. How can I predict who will terminate among current employees?

Termination reason would be important too (three categories that seem to matter are:

* termination/departure within a week or two of hire

* voluntary termination (e.g. quit for another job)

* getting fired

)

Importance: Answering this question will help us analyze the difference between different jobs/departments and come up with specific retention plans for those different jobs

Variables tracked: Hire_Date, Term_Date , Job, Hourly Rate, Full/Part time, Shift

Variables not tracked:Source of hiring, The time the employee took the offer, History of long-term successful employment, Reason for leaving the last job

Methods of analysis:

Nice to see some illustrations of methods here!

i. Logistic regression: Termed?` ~ Merged.Job + Hourly.Rate + Full.Part.Time + Shift

if shift turns out to be significant, what action would it make sense for the company to take? they have to hire for that shift anyway....

(I'm not saying don't include it; just be thoughtful about how you interpret it for the company!)

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    11.64754    1.23238   9.451  < 2e-16 ***
Merged.JobOther                -2.96472    0.46419  -6.387 1.69e-10 ***
Merged.JobWarehouse - Other    -2.31818    0.48877  -4.743 2.11e-06 ***
Merged.JobWarehouse - Selector -0.82588    0.38418  -2.150   0.0316 *
Hourly.Rate                    -0.55628    0.05623  -9.893  < 2e-16 ***
Full.Part.TimeP                -0.57837    0.53930  -1.072   0.2835
ShiftN                          0.50806    0.29529   1.721   0.0853 .
---
```

1. Higher Hourly rate, Part time job, day shift jobs are more likely to stay at the company (Termination = No). People in Other and

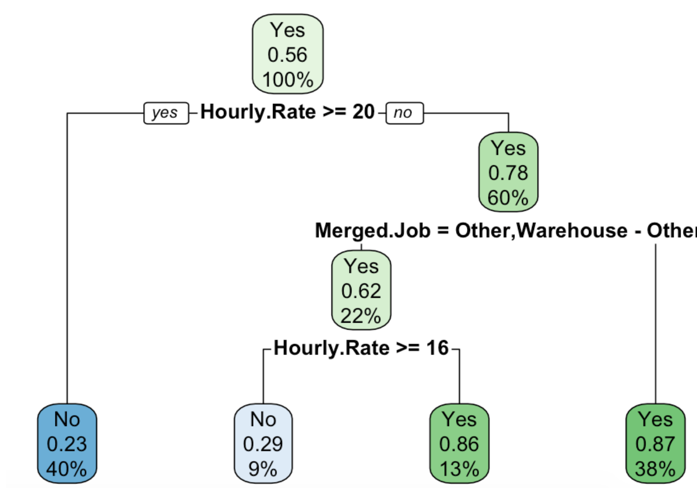
```
          Reference
Prediction No Yes
       No  60  16
       Yes 10  60

              Accuracy : 0.8219
```

I don't understand this table.
Does "reference" = "truth"?

Both the logistic regression and the tree model are nice to see here.

   Warehouse -other positions are more likely to stay.
2. With a 0.6 threshold, our model has an 82% accuracy. For the 146 observations (people) used in the test set, the model correctly predicted whether or not somebody terminated 82% of the time

ii. Tree model:



1. People who are more likely to stay at the company are those who have an hourly rate > =20, and people in Other and Warehouse-other departments with an hourly rate between 16 and 20.

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  56  17
       Yes 14  59

               Accuracy : 0.7877
```

2. Our model has a 78% accuracy. For the 146 observations (people) used in the test set, the model correctly predicted whether or not somebody terminated 78% of the time.

I think they will like this

7. **What types of employees are more likely to stay at the company for more than 6 months?**

Importance: Understanding the difference of people staying at their jobs will help the company design specific hiring strategies for those departments. If "other" jobs (non warehouse/ drivers) have a higher leaving rate, then we know that the problem is not unique to the warehouse and transportation department.

employee demographics will matter here too

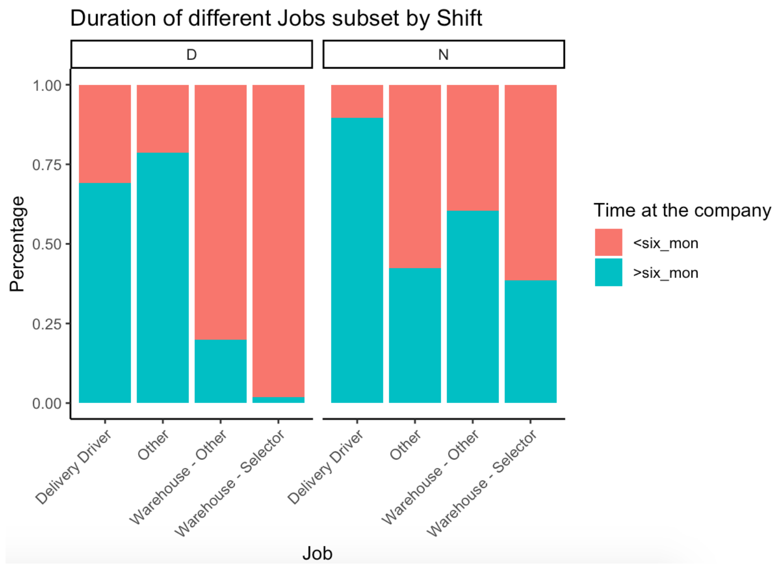
also things like "last job held", "amount of time at last job", etc.

Variables tracked: Job , full/ part time, Day/Night Shift

Variables not tracked: Source of hiring, Offer acceptance rate , Reason of termination, Employee satisfaction, either through surveys or websites like Glassdoor.com

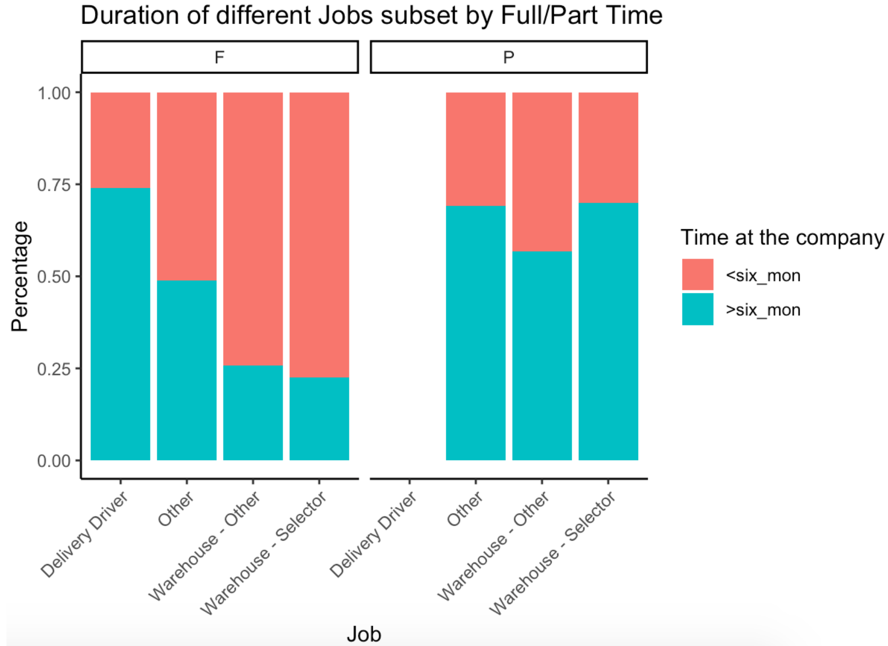
can you link to specific employees?

Methods of analysis:



1. From this plot, we can see that for Day shift jobs, there are more people in delivery drivers and other job positions that stay > 6 months. For night shift jobs, there are more people in delivery drivers and warehouse other positions stay > 6 months. Comparatively, we see that the warehouse and transportation positions usually stay more than 6 months in the night shifts but not for other job positions. It is probably because the pay rate is usually high for Night shifts and in order to further analyze our assumption, we need to do a survey on reasons of termination.

is this actually true or just guessing? want to rephrase slightly if just guessing.

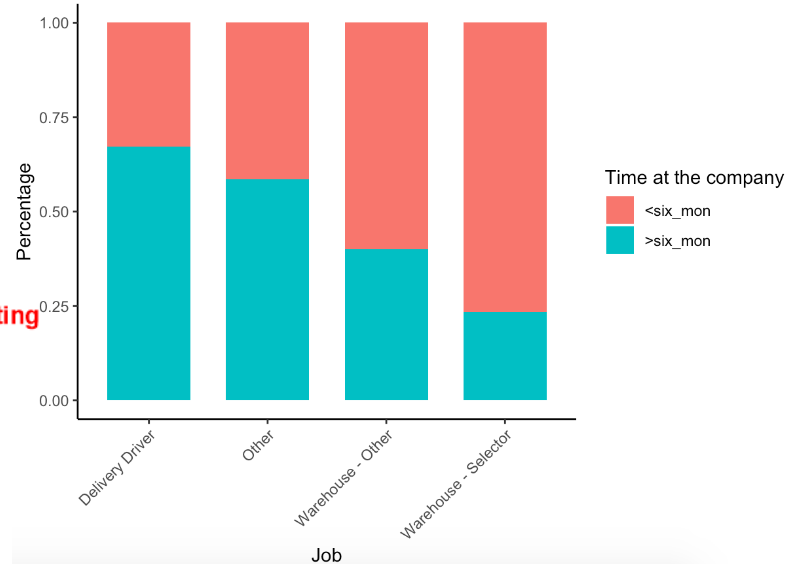


2. Another factor we could analyze based on our current data sets is the full and part time. From this plot, we could see that for full time positions, people in delivery driver positions have the highest rate of staying > 6 months, even higher than “other” positions. For warehouse positions, we could see that they usually stay more than 6 months as a part time job. Comparatively, we see that people in warehouse positions would stay long as a part time job but not as a full time job. The “reasons of termination ” will help us better analyze why such difference exists.

is this the same as the graph on the left above?

or is it aggregated some-how?

make clear before presenting graph.



3. This is the overall percentage of people who stay more than 6 months in our merged data set. We see that indeed, transportation positions have a higher rate(similar rate as) than “other” jobs, so we should focus more on the warehouse selector jobs

8. **Are employees who stay more than 6 months more likely to stay more than a year?**

No big jump (not sure if we need this questions)

bummer!

But another way to look at it would be as a suggestion that the company should find ways to create a jump.

Candidate attraction

9. **What are the common traits of hired applicants?**

great

Importance: Traits exhibited before the screening process such as time takes to apply the position, application source and traits exhibited during the screening process such as time to get hired may have a correlation with how applicants will react to the job once they get hired. With this information, one can have a general idea on the likelihood of the applicant accepting the offer or performing satisfactorily in the job. Assumption: “Successfully hired applicants” is defined as hired applicants that accept the offer, stay in the company for more than one month and with satisfactory performance.

Variables tracked: Application Date, Division, Requisition Open Date, # of Openings, Status

Variables Not Tracked: Source of Hiring, Link between application ID & employee ID, Employee Performance Rating

Method of Analysis

General comment: in final report/presentation, give examples of using the methods for each question to help answer (or get closer to an answer for) each action question.

i. Group hired applicants based on time to apply the position, hire source and time to get hired.
ii. For each group of applicants, calculate their rate of acceptance.
iii. For each group of applicants, calculate their distribution of performance ratings.

10. How can we attract more candidates to apply?

great

Importance: We can see that for driver and warehouse positions, especially driver positions, our clients are facing recruitment issues, hard to fill the expected headcount. If we can attract more candidates to apply, at least a part of the recruitment issue should be addressed.

Variables tracked: Job descriptions (this can help us determine whether clear information is presented. If not, some candidates that could apply may miss the chance. Or, some candidates who apply will finally find out he or she does not fit this position, and it will cause a waste of time for both the company and the candidate)

Variables Not tracked

i. Source of application of all candidates (if we can now the source of application, we can see the proportion of candidates of each source, and determine which source to focus more on. Say, if most of the candidates apply for the job through Linkedin, the company may should post and update the job description more frequently. Also, the HR can reach potential candidates through Linkedin.

ii. The number of hit on the company website job page, including the hit on different job positions (This is also a source that we can track)

Method of analysis

General comment: in final report/presentation, give examples of using the methods for each question to help answer (or get closer to an answer for) each action question.

iii. We can build a pie chart and a bar chart to visualize the proportion of candidates from each source. And also we can build a time series plot for each source to see if there is any change by time. Thus, we can determine which source to market and to focus on so that we can attract candidates. Say, one of the clients said that they used to attend some career fairs in person. However, only a few people came to the career fair nowadays, and they decided not to attend those events now.

Descriptive Questions:

1. What is the average time to hire for each job role?
2. Is there seasonality in hiring and termination?
3. Are there difficulties in filling requisitions (does the status change from time to time, region to region)?
4. What are the qualities of the application pool from the recruiting point?
5. What is the number of qualified applicants per hire?
6. What is the average time between application date and the requisition opening date for the __________ job role?
7. What is the ratio between complete and incomplete applications? (Application Completion Rate)
8. How many applicants do you receive per opening?
9. What is the target group for ______ job role?
10. What is the average length of each stage of the interview process?
11. Why did the candidate leave the previous employer?
12. For new hires, what is the drop-off rate within the first two weeks? (including the window in which the most drop-off occurs in)
13. What is the offer acceptance rate?

Web Analytics Questions:

Web analytics are helpful to understand the conversion from awareness to interest, as well as from interest to application. By analyzing web-specific activity data, the recruitment team can have more perspectives on the digital recruiting landscape and increase the number of applicants. We address the concept of Awareness in the recruiting pipeline to quantify how many viewers are looking at the job postings online through the various sourcing channels, and mainly the clients' careers page. By observing how many views these job postings are receiving, we can derive the Interest in these roles as viewers who start the application process, whether it is complete or incomplete. We can identify gaps between Interested candidates and completed applications by observing the pageviews or web-traffic within each page of the application process online.

- **What proportion of visitors access the company website careers page using each device (mobile / desktop tablet)?**
  - Importance: User experience from different devices usually vary a lot. If there is a considerable amount of users accessing the company recruitment page using mobile devices, the company can consider developing a mobile-friendly layout.
- **What is the conversion rate of each web sourcing channel?**
  - Variable definition
    - Interest conversion rate = (number of visitors who clicked the "apply" button) / (number of unique visitors)
      - This is essentially the apply button "click-through rate".
    - Application conversion rate = (number of applicants who completed application) / (number of visitors who clicked the "apply" button)
- **What are the sources of the visitors on the website careers page?**
  - Importance: By tracking the number of visitors from different sources, the company can have a more holistic view of the effectiveness of sourcing channels.
  - Traffic sources examples
    - Social media
    - Job boards
    - Paid advertisement
    - Direct traffic (website visits that arrived on your site either by typing your website URL into a browser or through browser bookmarks)
    - Organic traffic (visits from search engines)