Manufacturing Vertical

Palantir

Assumptions - Manufacturing

- Simple random sample of 200,000 job postings used for analyses. (Original dataset has 8M rows)

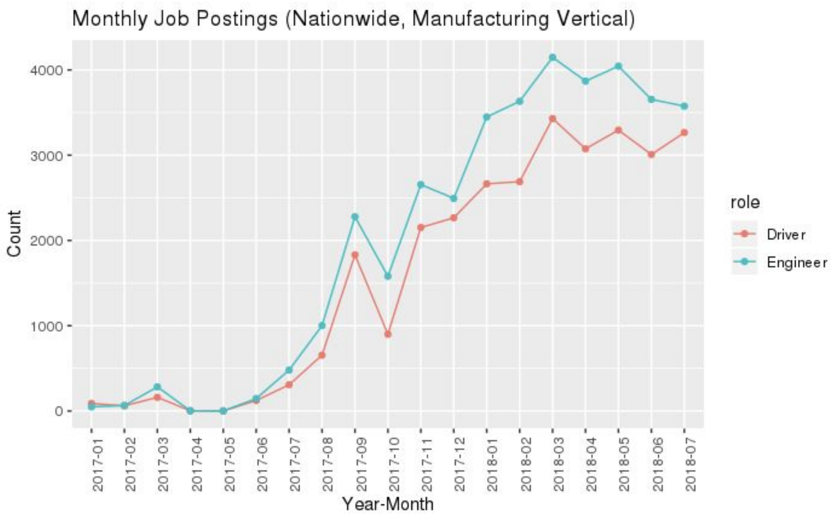
- Analyses used job postings August 2017 - July 2018.

- time_to_fill under 3 days or over 123 days ignored.

- NA rate for time_to_fill was ~2.5%.

- Combined “Engineer” roles (*n* = 41,899) and “Driver” roles (*n* = 33,796) due to similar time_to_fill and salary characteristics (could be expanded back out).

- Selected engineers and drivers as our two roles - most frequent postings, simplify the vertical.



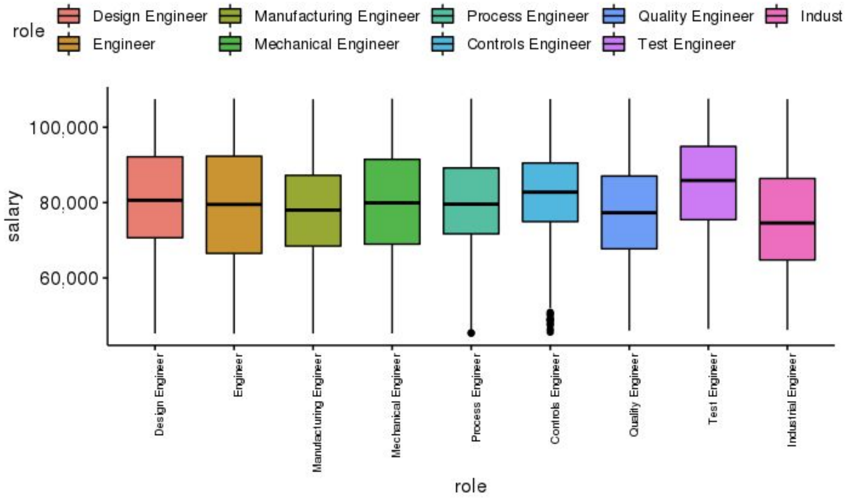
EDA: Manufacturing - Engineers

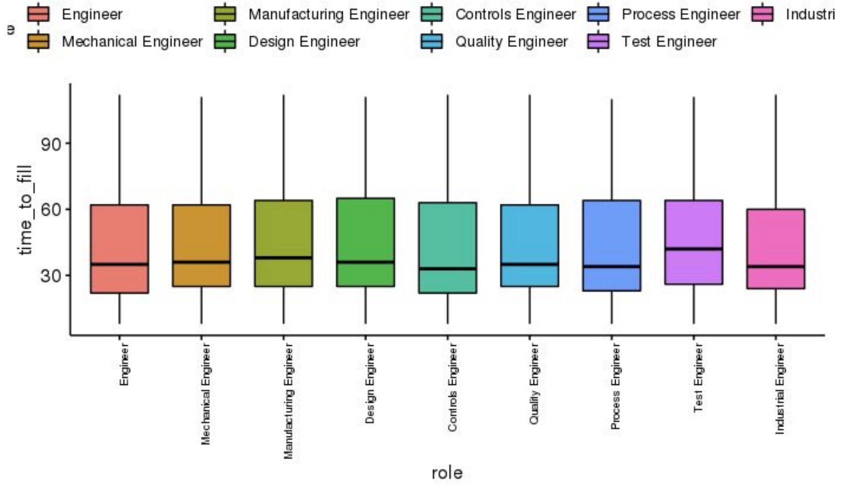
Overview of Data - Combining Roles

- Took top 99% of engineer roles (by postings) and combined into one group for modeling and analysis.
- Removed bottom 10% and top 10% of salary.
- Time_to_fill in days:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.00	19.00	35.00	47.69	72.00	122.00

- For the top nine engineer roles, top tag counts (“Electrical”, “Automation”, “Safety”, etc…) were aggregated into vectors and tag counts were compared using Euclidean distances. All 45 pairwise distances ranged from 1.277 to 2.776, which are fairly close. So all engineering roles were combined into one “engineering” group.





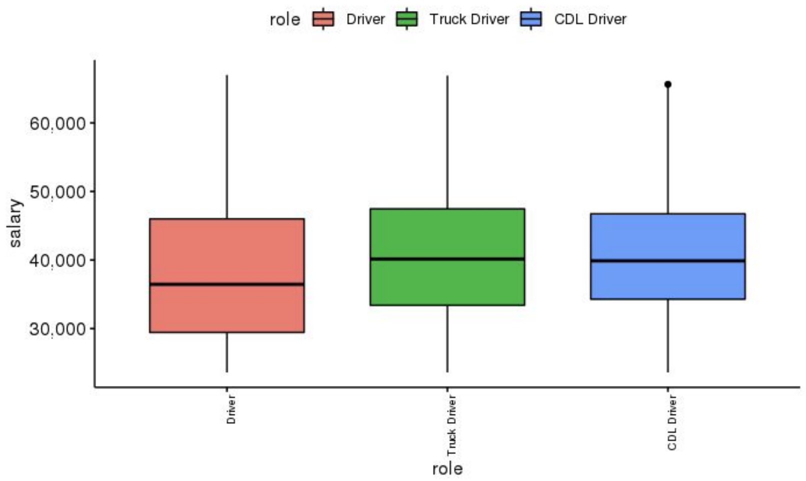
EDA: Manufacturing - Drivers

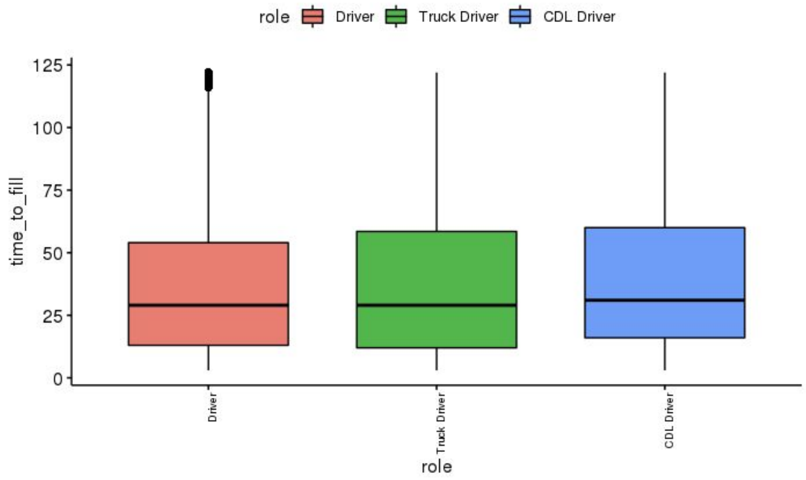
Overview of Data - Combining Roles

- Removed bottom 10% and top 10% of salary (dozens of salaries exceeding $100,000 or even $200,000).
- Time_to_fill in days:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.00	13.00	29.00	38.56	54.00	122.00

- For the three driver roles, top tag counts (“Safety”, “Commercial”, “Transportation”, etc…) were aggregated into vectors and tag counts were compared using Euclidean distances. All three pairwise distances were 1.676 or below and the roles were combined into one general “driver” group.





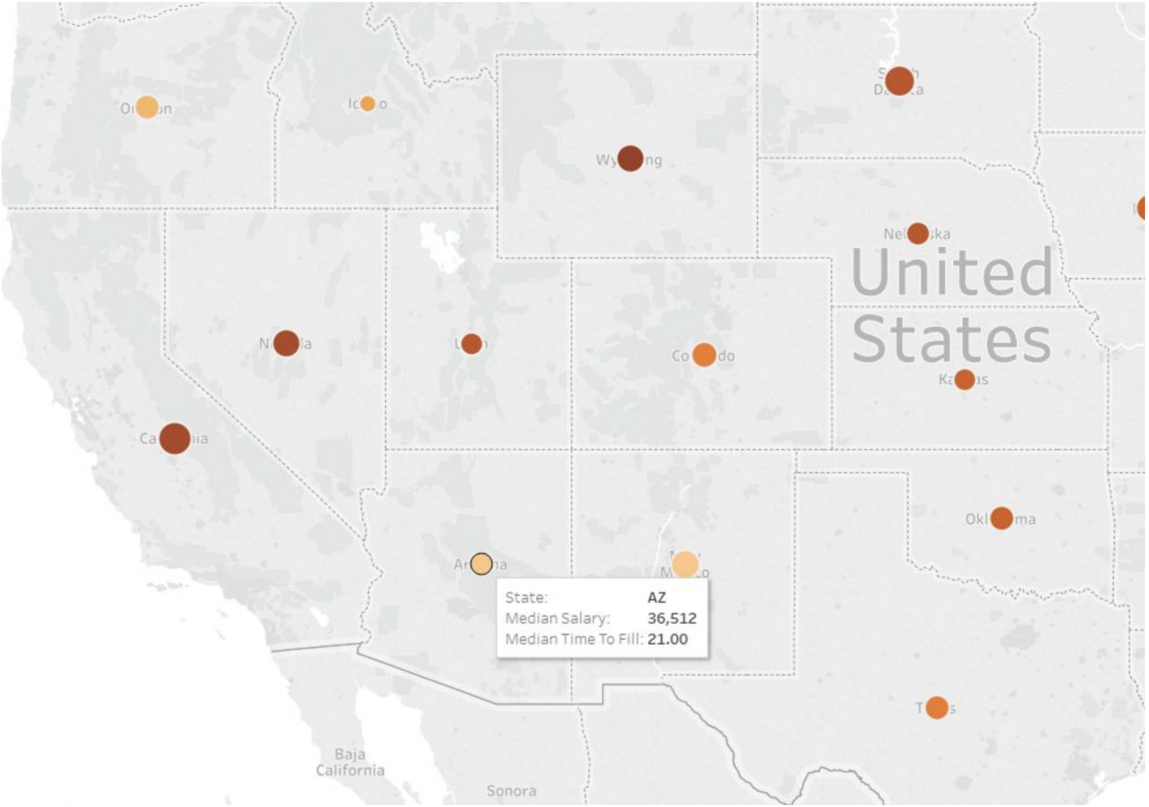
EDA: Manufacturing - Engineers and Drivers

- State level summaries are available but ideally we would break this down to metro_state instead.

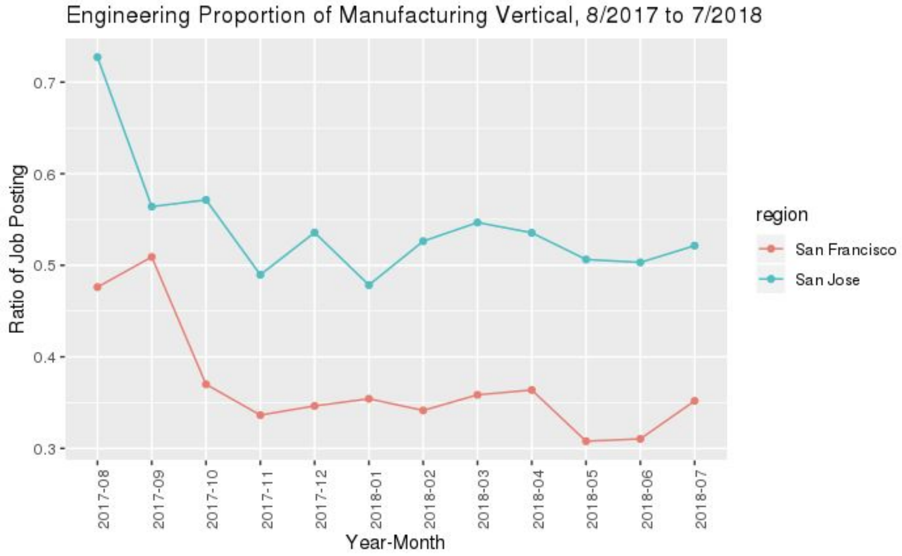
- To the right, median salaries and time_to_fill for drivers.

    - Darker circles have longer time_to_fill.

        - Larger circles have larger median pay.



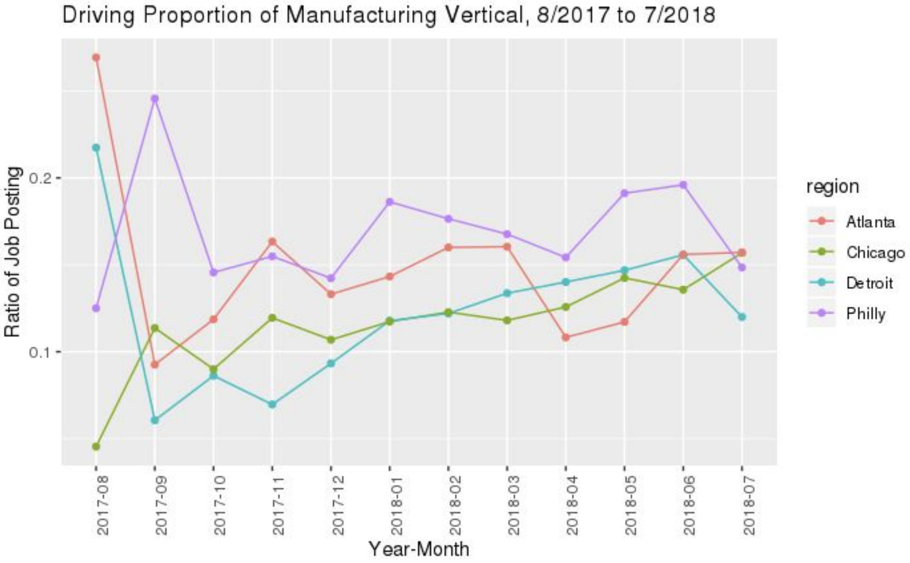
Manufacturing Trendlines

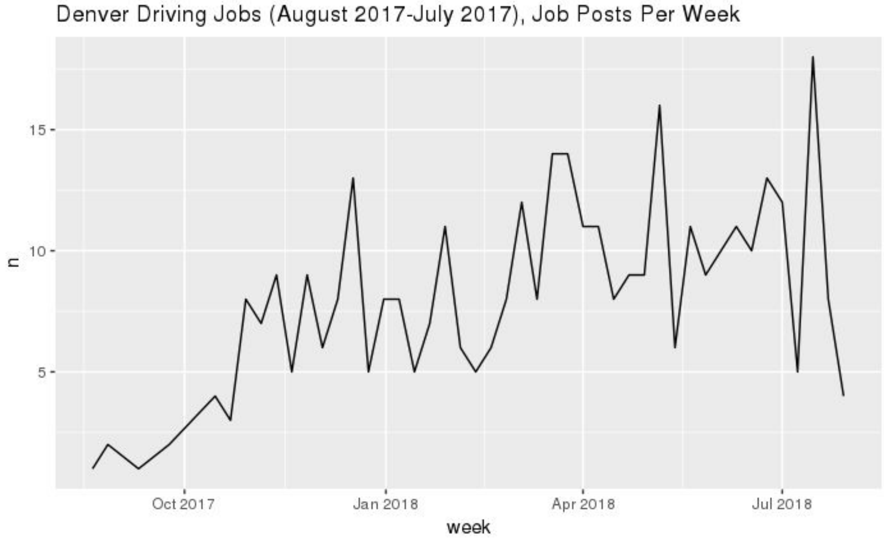


Left: Job postings in each role as a proportion of total manufacturing job postings

Right: Two select markets, engineering postings as a proportion of region_state manufacturing job postings

Manufacturing Trendlines

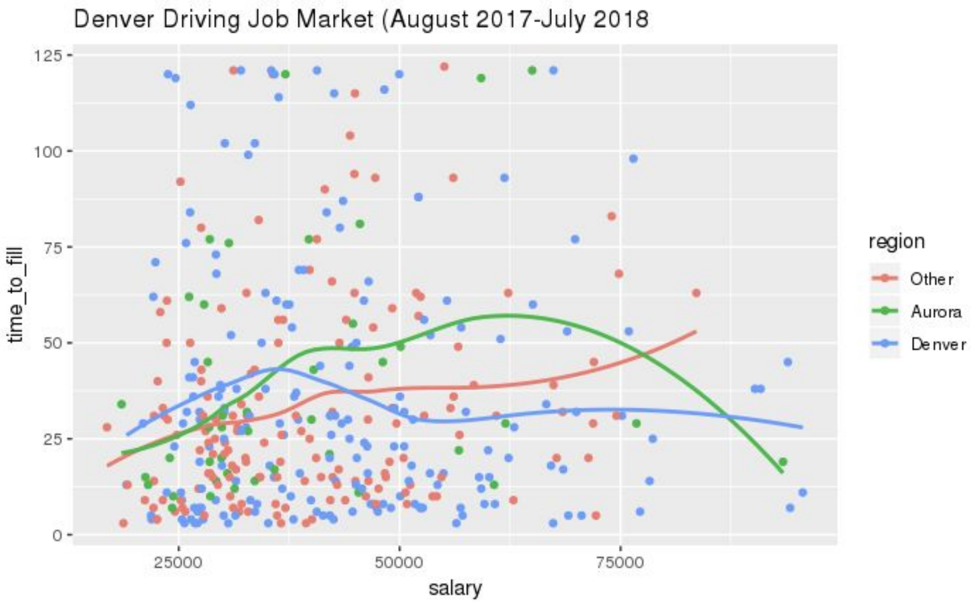




Left: Four select markets, driving postings as a proportion of region_state manufacturing job postings

Right: Denver driving jobs, number of postings over time - short predictive model on time_to_fill

Predicting Time_To_Fill: One Job, One Market

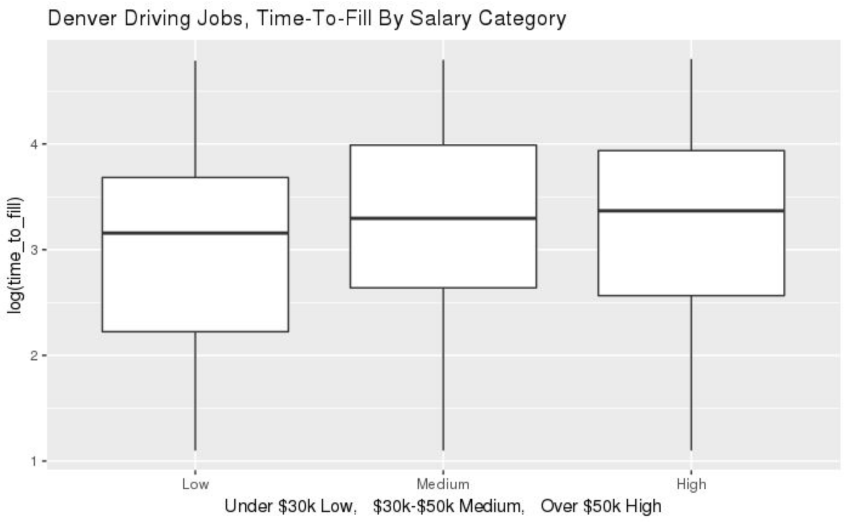


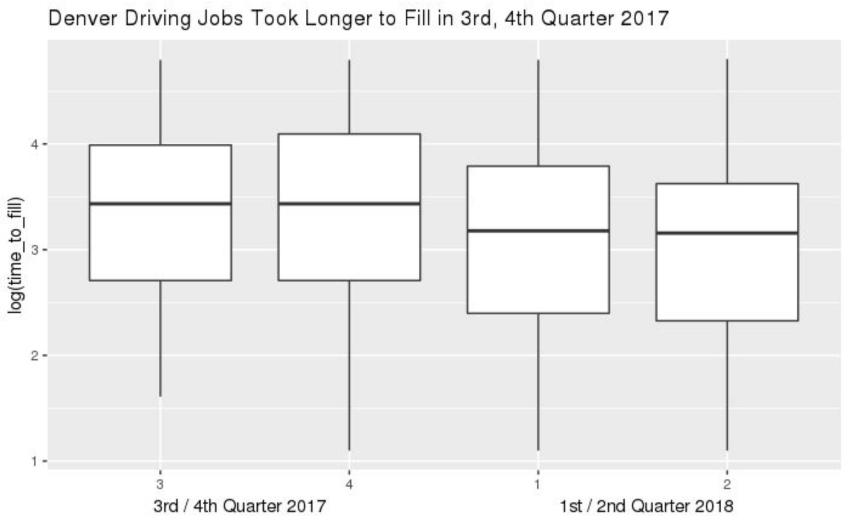
- Split Denver metro into three regions: Denver, Aurora, and suburbs
- Relationship between salary and time_to_fill is weak, added other features:
- From tag analysis, most frequent driving tags are “Safety”, “High School Diploma”, “Training”, “Transportation”, “Customer Service”, “Commercial”, and “Warehouse”
- Created a quarter indicator: “Q3 2017”, “Q4 2017”, “Q1 2018”, “Q2 2018”.
- Created salary buckets with ~equal job counts: “Low” (under $30k), “Medium” (up to $50k), “High” (over $50k)
- Ran a multiple regression with best subsets using AIC but didn’t find much predictive value.

Predicting Time_To_Fill: One Job, One Market

- Although the overall model was reasonably significant, P-value = 0.0111, doubts linger on extending this type of analysis to other markets or a broader geographical region (Adj R-sq = 0.0217).

  Time_to_fill = 1.09 + 0.21(*log salary*) + 0.17(“*Transportation*”) - 0.19(“*Warehouse*”) - 0.20(*Q1*) - 0.29(*Q2*)



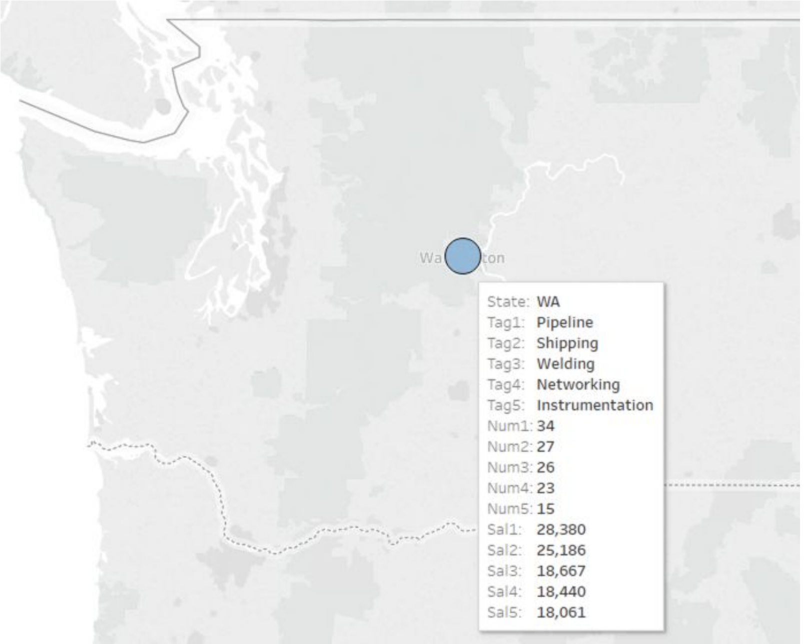


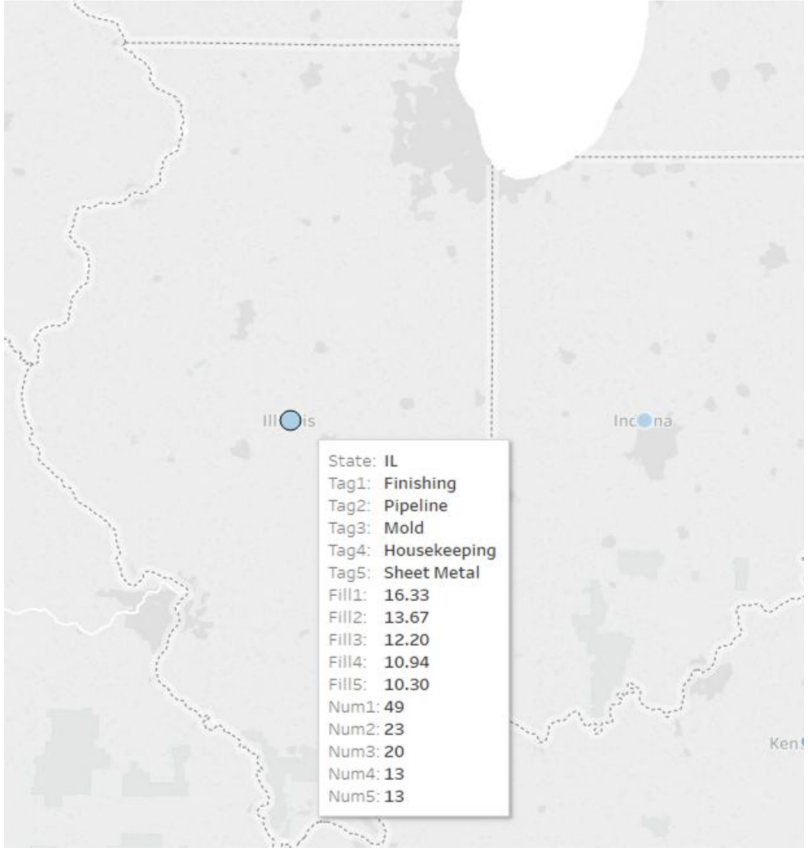
Analyzing Tags - Engineers (Nationwide)

Tags	Percent of Postings	Dollars per Occurrence
Pipeline	10.50%	$20,266.82
Masters Degree	10.14%	$17,713.16
Mold	9.78%	$15,674.05
Printing	10.24%	$15,474.69
Fleet	9.67%	$15,267.21
Transmission	9.63%	$15,173.79
PhD	7.79%	$14,876.68

- Top seven engineering tags by total dollars / job posting.
- Could be reduced to individual metro regions or fine-tuned to specific engineering roles.
- Potential value in knowing which tags are contributing towards salary.

Engineering Tags





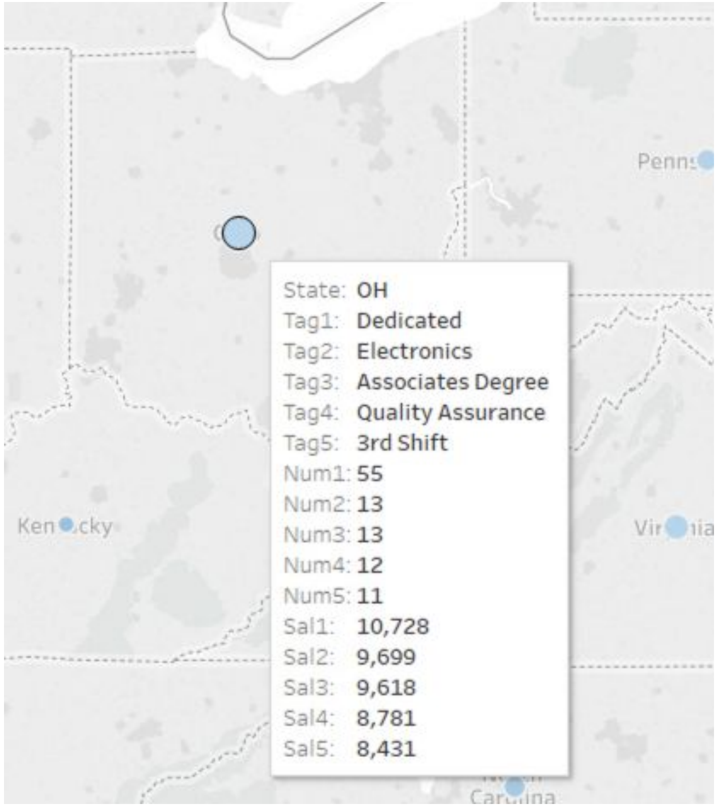
Left: By state, top five most frequently occurring tags, dollar value per job with tag
Right: By state, top five longest-to-fill tags, in terms of mean day

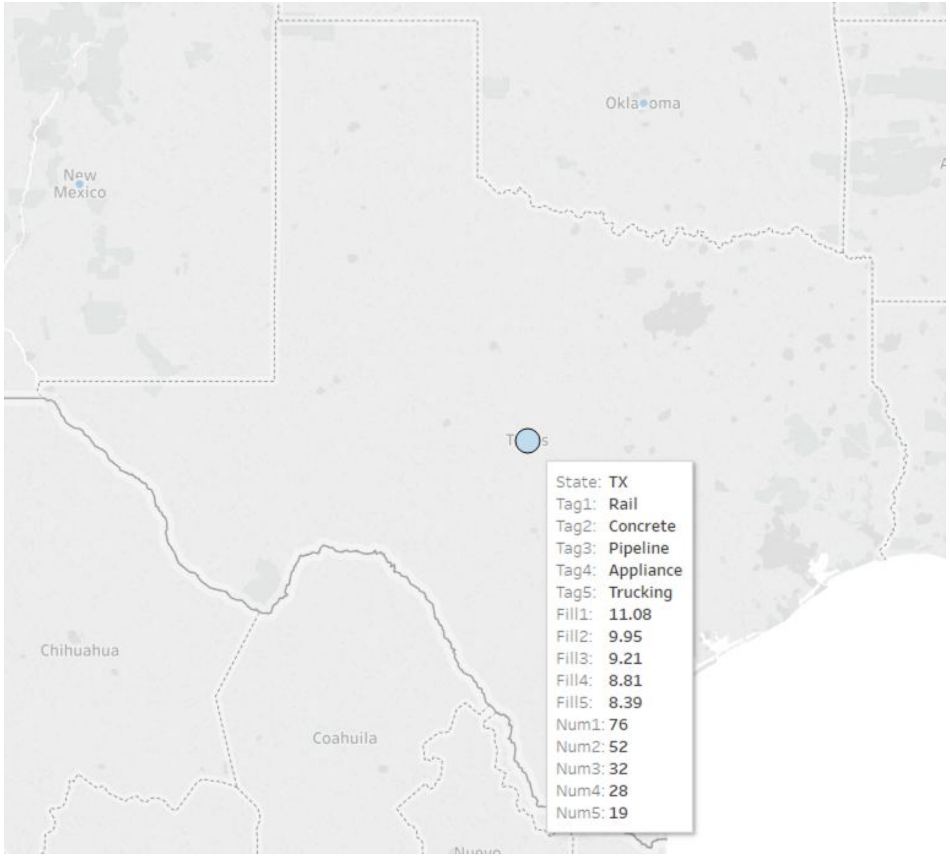
Analyzing Tags - Drivers (Nationwide)

Tags	Percent of Postings	Dollars per Occurrence
1st Shift	7.14%	$8,088.05
Finance	6.23%	$7,869.40
Dedicated	6.09%	$7,865.56
Programming	9.13%*	$7,791.14
2nd Shift	7.96%	$7,778.47
Crane	8.54%	$7,764.74
Trucking	7.57%	$7,761.91

- Top seven driving tags by total dollars / job posting.
- “Finance” and “Programming” potential red-flags*.
- Potential value in knowing which tags are contributing towards salary.
- Extendible to other roles, or reducible down to specific geographic regions or time periods.

Driving Tags





For example, in Ohio, “3rd shift” tagged jobs contribute $8,431 to salary.

Predictive Models For Engineers (Nationwide)

- Predicting time_to_fill within one month / over one month
- Model predictors are state, month, and salary
- 80% / 20% split on training / test data
- Same assumptions as before (time_to_fill under 3 days, over 123 days removed, use central 80% of salaries, etc..)

Random Forest

	actual	
pred	< 1 month	> 1 month
< 1 month	2075	1593
> 1 month	877	1274

Correct Class Rate: 57.55%
Precision: 59.22%
Recall: 44.44%
F-Score: 0.5078

Support Vector Machine

	actual	
pred	< 1 month	> 1 month
< 1 month	2370	1932
> 1 month	582	935

Correct Class Rate: 56.80%
Precision: 61.64%
Recall: 32.61%
F-Score: 0.4266

Predictive Models For Drivers (Nationwide)

Random Forest

actual		
pred	< 1 month	> 1 month
< 1 month	2873	1370
> 1 month	97	336

Correct Class Rate: 68.63%
Precision: 77.60%
Recall: 19.70%
F-Score: 0.3142

Support Vector Machine

actual		
pred	< 1 month	> 1 month
< 1 month	2933	1441
> 1 month	37	265

Correct Class Rate: 68.39%
Precision: 87.75%
Recall: 15.53%
F-Score: 0.2639

Conclusions:

- Methodology (R.F. or S.V.M.) is inconsequential as performance measures are close
- Engineering models are only slightly beating a random coin flip.
- For drivers, precision is relatively high (of the predicted 1+ month_to_fill, a high % are correct)
- For drivers, recall is very low (of the true 1+ month_to_fill, model is only predicting a low %)