MSiA 400 Team Palantir

Greenwich Project

Anjali Verma

Joseph Kupresanin

Lingjun Chen

Yueying Zhang

Ziying Wang

1. Introduction

In this project, we analyzed the labor demand in manufacturing area using the job posting data provided by Greenwich. Our analysis mainly focused on the Engineer and Driver jobs since they have the largest job postings in manufacturing. Our report is outlined as below:

In section 2, we covered the necessary data cleaning and exploratory data analysis to make sense of different aspects of the data. Specifically, we found strong relationship between demand and salary, demand and time to fill. As the result, understanding salary and time to fill can help form a clearer picture of the labor demand in manufacturing. Also, we explored how labor demand changed over time and across different markets.

In section 3, we analyzed how skills shape factors related to demand such as salary and time to fill for Engineer and Driver jobs. We identified the skills that are valued most in terms of the additional earning they bring and the difficulty of finding people meeting those skill requirements in labor demand market.

In section 4, we used both Random Forest, Support Vector Machine as well as logistic regression to classify time to fill into "within one month" or "over one month". We believe companies approaching Greenwich will be interested in how long the position can be filled, and therefore it will be valuable to Greenwich if they can have such a prediction model. In addition, we did a mini case study of the Denver Driver market. We built a linear regression model to predict transformed time to fill. We found time and the appearance of certain tags are significant predictors.

In section 5, we concluded our findings and addressed the potential improvement we can make.

2. Data Cleaning and Exploratory Data Analysis

2.1 Data Source

We are provided with three datasets by Greenwich which are greenwich_master, greenwich_role and greenwich_tags. The greenwich_master contains the basic information of the posted position such as estimated salary and location. The greenwich_role indicates the job responsibility, and the greenwich_tag specifies the skill requirement for each position. We randomly selected 200,000 job ids out of original 8M rows for our analysis.

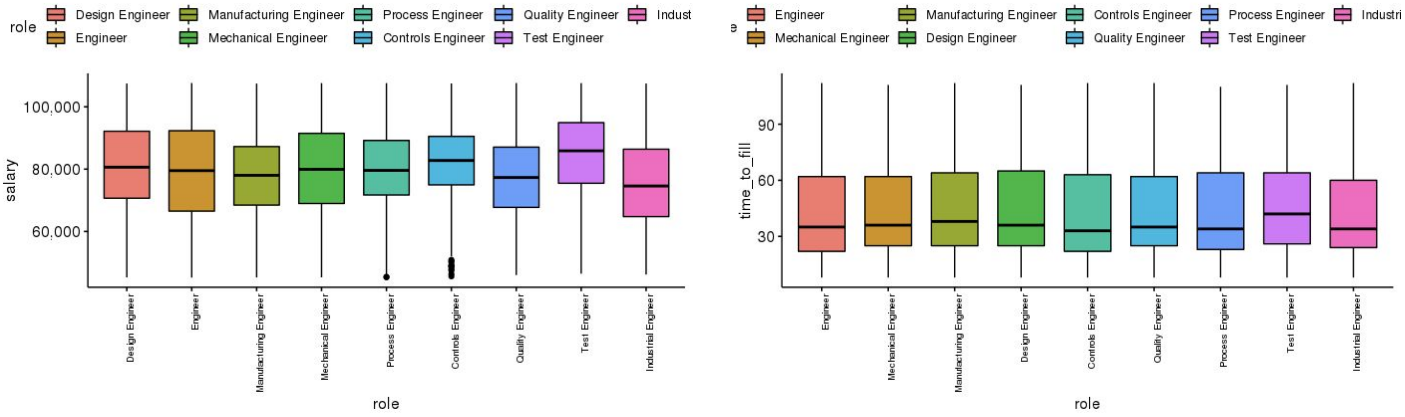
2.2 Data Cleaning

Some companies post job online only for internal requirement instead of truly looking for labor. The time to fill for this type of posting is very short. Therefore, we ignored the job postings whose time to fill is under 3 days and also for those that > 123 days. We also removed the job posting having salary in the bottom and top 10% in the salary distribution. Some of those jobs have salary exceeding $200,000, which can be outliers and cause bias in our analysis results.

2.3 Exploratory Data Analysis

We found for different sub-types of Engineers such as Process Engineer and Controls Engineer, their distributions of salary and the distributions of time to fill are similar. Therefore, we combined all Engineer jobs into one role called Engineer for further analysis and modeling. We

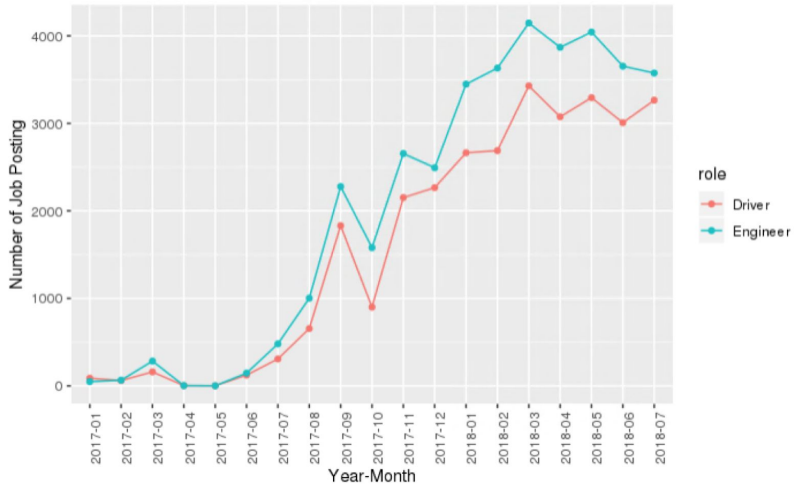
also aggregated all Driver jobs (Truck Driver and CDL Driver) into one single role called Driver based on their similar distribution of salary and time to fill. After combination, we have approximately 42,000 Engineer jobs and 34,000 Driver jobs.



We found there are significant and negative correlation between time to fill and demand, significant and positive correlation between salary and demand. As a result, predicting the change in time to fill and salary will help us better understand the labor demand.

2.4 Demand for Engineer and Driver

The first metric to measure the demand for Engineer and Driver is number of job posting. The job posting is summed up by month. The time series graph shows number of Engineer and Driver job posting displays an increasing demand for the two roles. However, one reason of this trend is that more data sources are added over 2017. Therefore, the number of job posting can only tell the relative demand change, but it fails to capture the demand change at absolute level.

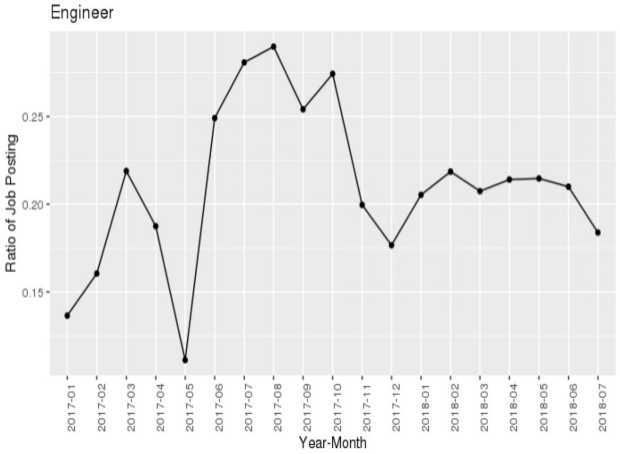


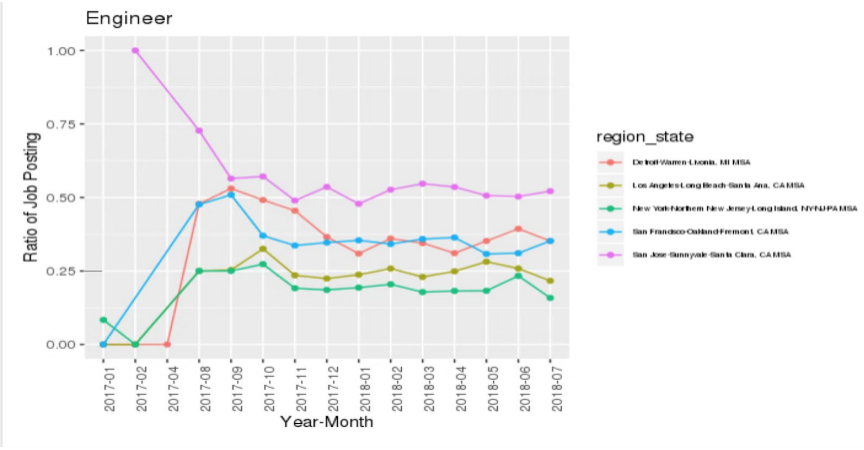
As a result, the ratio of job posting is considered and utilized in the further demand analysis. The ratio of job posting of Engineer is calculated as the job posting of Engineer role divided by the total job posting of all roles. Same definition is also applied in defining the ratio of job posting of Driver. We believe it is reasonable to make the assumption that with large volume of data, the proportion of job posting of each role in the added data sources is stable.

Between Jan, 2017 and May, 2017, the volume of data is small. One unit increase in Engineer or Driver role will have strong impact on the ratio, and therefore this time period is not considered in the analysis. The demand for Engineer drops from Oct, 2017 and becomes stable since Jan, 2018. Similar trend can also be observed in the top 5 region states with highest number of job posting of Engineer. The top 5 regions state are :

1	Los Angeles-Long Beach-Santa Ana, CA
2	San Francisco-Oakland-Fremont, CA
3	New York-Northern New Jersey-Long Island, NY-NJ-PA
4	San Jose-Sunnyvale-Santa Clara, CA
5	Detroit-Warren-Livonia, MI

California has the highest demand for Engineer over time. In Oct, 2017, all top 5 regions face a drop in Engineer demand.

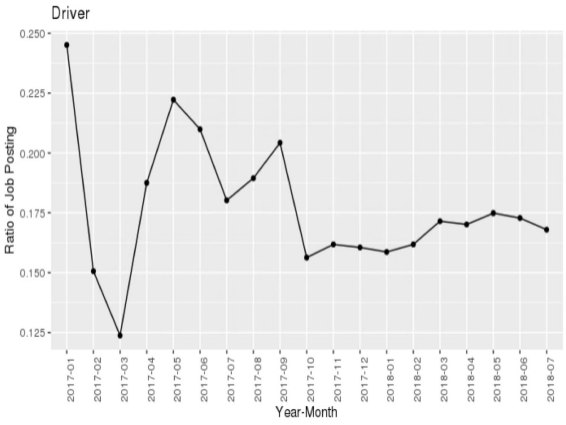


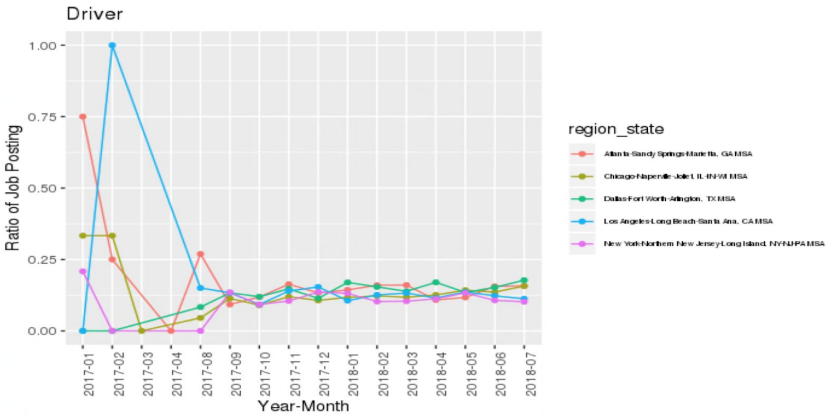


The demand for Driver has a spike in Sep, 2017 and becomes stable since Oct, 2017. The top 5 region states with highest demand for Driver are:

1	New York-Northern New Jersey-Long Island, NY-NJ-PA
2	Los Angeles-Long Beach-Santa Ana, CA
3	Chicago-Naperville-Joliet, IL-IN-WI
4	Dallas-Fort Worth-Arlington, TX
5	Atlanta-Sandy Springs-Marietta, GA

Compared with the demand for Engineer, the demand for Driver is less volatile and does not vary a lot among top region states.





3. Skill Analysis

We looked at how tags (skills) affect demand of Engineer and Driver by analyzing what the top wanted skills are and how each tag contributes to key metrics for demand such as salary and time to fill.

3.1 Most Common Skills

For Engineers jobs, the top 7 skills with their counts are:

1	2	3	4	5	6	7
Lead	Excel	Planning	Electrical	Travel	Safety	Industrial
8287	6791	6394	6322	6296	6270	5994

Engineering jobs require specific skills such as Electrical and Industrial. The top 1 tag for engineers is lead, indicating that the industry wants engineers to be able to lead their projects.

3.2 Most Valuable Skills

We had 599 tags appearing in engineer jobs originally. For our further analysis, we excluded the tags appearing less than 1% of job postings since the least frequent tags may only represent the special requirement of few companies.

Next, we used the formula on the right side to calculate the contribution of each tag to the salary of one job posting. The Tf-idf value shows a relative frequency of the occurrence of tags in all the postings. The larger the value, the less frequently the tag appears. The less frequent tags will be distributed with a larger portion of the salary since they are more likely to be the valuable and preferred qualifications instead of the basic ones. However, only looking at the sum of the earnings for each tag can be biased since a tag appearing more times can have larger sum even though it has a smaller Tf-idf value. Therefore, we regularized each tags earning by the number of time it appears and identified the top 7 skills with largest earning per occurrence.

$$\text{for each job posting } i:$$
$$\text{assume it has } m \text{ tags}$$
$$tag_j \text{ earning} = salary_i * \frac{Tfidf(tag_j)}{\sum_{i=1}^{m} Tfidf(tag_i)}$$

Rank	1	2	3	4	5	6	7
Tag	Pipeline	Masters Degree	Mold	Printing	Transmission	PhD	Fleet
Dollars per occurrence	19592	17657	16005	15470	15352	14725	14697
Percentage of posting	10.50%	10.14%	9.78%	10.24%	9.67%	9.63%	7.79%

On average, the salary will be increased by $19,592 if having the knowledge of pipeline appears in the job requirement. In addition, having more advanced degrees will result in higher salary.

3.3 Most Demanding Skills

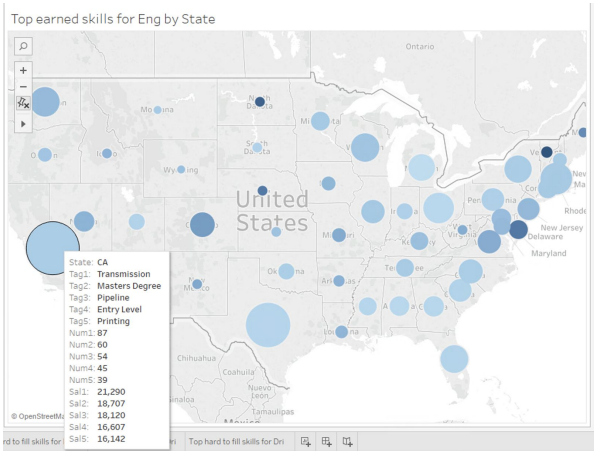
Using the similar Tf-idf methods as above, we identified the top 7 skills associated with longest time to fill.

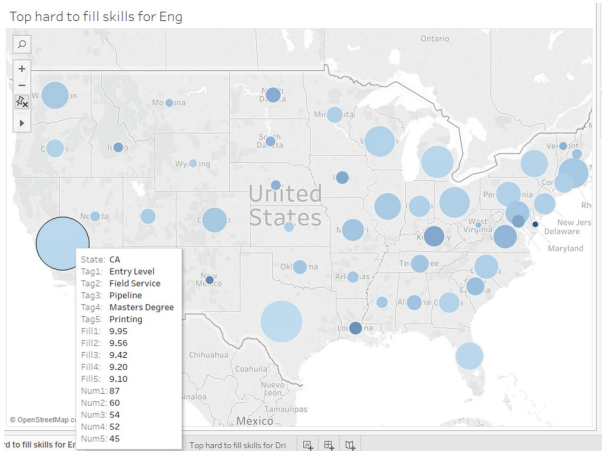
Rank	1	2	3	4	5	6	7
Tag	Masters Degree	Mold	Printing	Pipeline	Transmission	Sheet Metal	Entry Level
Time to fill	8.89	8.71	8.55	8.53	8.09	7.98	7.97

If having a master degree appears in the job requirement, the average time to fill the job will increase by 8.89 days. We found that the tags earning the most per occurrence are strongly correlated with those having longest time to fill with a 0.93 correlation. This makes sense since employers are willing to pay more for skills that are harder to find, which shows how these skills shaped demand in the industry.

3.4 Skills across States

We take a step further to look at the skills having highest monetary return and longest time to fill for different states. The Tf-idf values were re-calculated within each state and used as weights to distribute either salary or time to fill. For California, the top 3 skills having highest earnings are Transmission, Masters Degree, Pipeline associated with approximately $20,000 additional earnings. The top skills having longest time to fill include Field Service and Pipeline with around 9 additional days to find right candidates.





We did the same analysis on Driver. The outputs are shown in Appendix.

4. Predicting "Time to Fill"

4.1 Predictive Model for Engineer (Nationwide)

In this section, we classified time to fill of Engineer and Driver jobs into either "within one month" or "over one month" using both Random Forest and SVM. Through analysis, we found time to fill varies across state, and it is influenced by seasonality. As a result, we included state, month and salary as our predictors. We used 80% of data as training set and the rest 20% as test set and performed a 10-fold cross validation. The average performances of both models are displayed below:

Engineer		
	Random Forest	SVM
Correct Classification Rate	57.55%	56.80%
Precision	59.22%	61.64%
Recall	44.44%	32.61%
F-score	0.51	0.43

Driver		
	Random Forest	SVM
Correct Classification Rate	68.63%	68.39%
Precision	77.60%	87.75%
Recall	19.70%	15.53%
F-score	0.32	0.26

Both Random Forest and SVM have similar performance measures. The model for Engineer only slightly beats a random coin flip. For drivers, precision is relatively high (a high percentage of the predicted "over one month" is indeed "over one month"). However, the recall is very low. The models tend to classify most samples into "within one month".

Additionally, we clubbed the driver and engineer jobs to run a logistic regression model with time to fill as the binary response variable and taking the same predictors as above. We took 0.4 as the cut-off parameter (that is the probability above which time to fill would be classified as within one month) since it maximised the correct classification rate at 66.26% as well as the F-score.

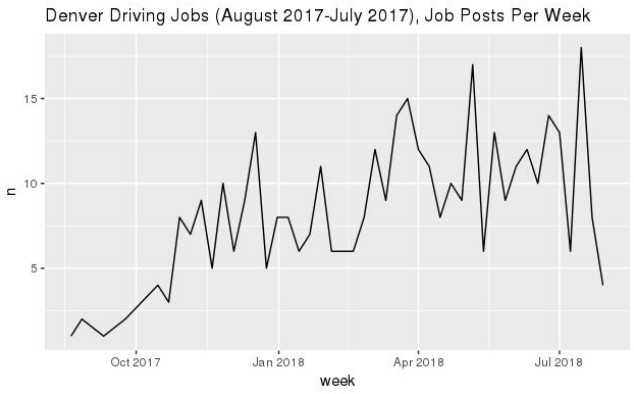
Engineers & Drivers	
	Logistic Regression

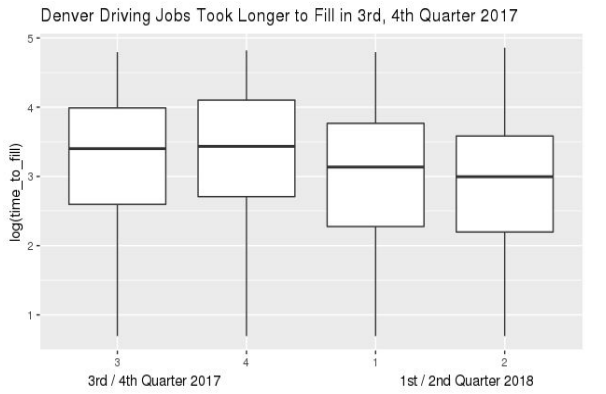
Correct Classification Rate	66.26%
Precision	53.62%
Recall	40.88%
F-score	0.46

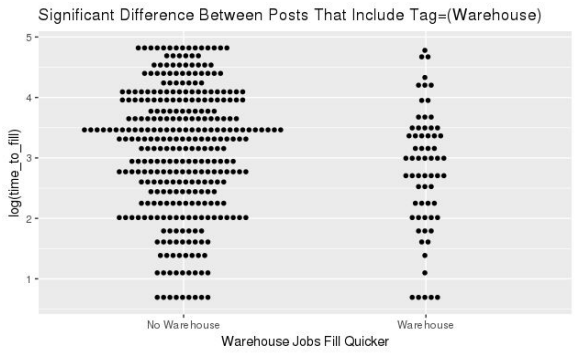
4.2 Predicting "Time To Fill" in Denver for Drivers: A Mini-Case Study

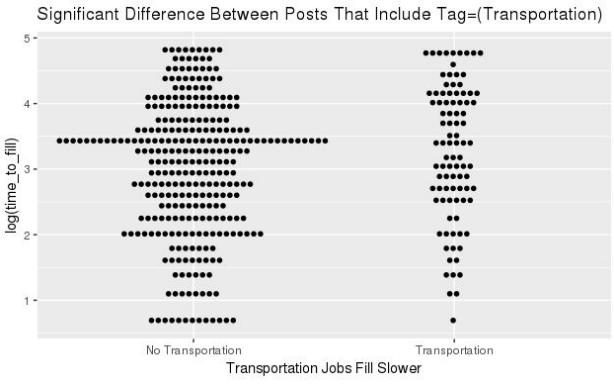
Because the U.S. job market is so large and varied (Seattle vs. Atlanta vs. rural Oklahoma / Occupational Therapists vs. Administrative Assistants vs. High School Science Teachers), one thought was to reduce to universe to one market and one job. This extreme variation in job characteristics will introduce significant noise into any analysis. Drivers in Denver were selected as a test case. If insight can be gained trying to predict "time to fill" here, the analysis could be extended to other markets and roles.

We found the market has an uptick in postings per week, but by mid Q1 in 2018, that trend seems to be leveling off. Minor but significant differences were found as shown below. Most significantly, on average, time to fill decreased in Q1 and Q2 for 2018 and jobs in Denver were filled more quickly. Tag="Transportation" increases the "Time To Fill", but tag="Warehouse" has the opposite effect. That indicates inter-predictor causality. Therefore, we decided to build a model with interaction terms such as time * tag.









We used “time to fill” with the Box-Cox transformation as dependent variable. This transformation on time to fill reduces the normality problem of linear regression. After implementing variable selection based on two-direction AIC, we had the model including normalized salary, popular tags of Driver, factorized quarter as well as the interaction terms with an adjusted R^2 0.11. The final multiple regression model is (shown in Summary 1 in appendix):

Transformed.time_to_fill ~ norm.salary + region + safety +diploma + transportation + cust.serv + training + commercial + warehouse + quarter + norm.salary*region + norm.salary*safety + norm.salary*diploma + norm.salary*cust.serv + norm.salary*commercial + norm.salary*quarter + safety*commercial + diploma*transportation + diploma*cust.serv + diploma*quarter + cust.serv*training + training*commercial + training*quarter + commercial*quarter

5. Conclusion

We used ratio of job counts to measure the labor demand of Engineer and Driver jobs. We found that regions in California have significant higher demand for Engineer. In the top regions with most job posting, we observed drops in Engineer demand from Oct, 2017. We also identified the top tags associated with highest earning and longest time to fill, which will give us insights on labor demand since salary and time to fill are correlated with demand. In addition, we built both classification and regression models on time to fill. Though the performances of the models are not ideal, we identified the factors that can impact time to fill such as the appearance of specific tags.

If time permitted, we would work on feature engineering to improve the performance of our classification and regression mode. For instance, we can cluster states into groups and use them as explanatory variables to better capture the relationship between location and time to fill.

6. Appendix

Table 1: Drivers Most Earning Tags

Rank	1	2	3	4	5	6	7
Tag	Shift	Dedicated	Programming	Landscaping	Finance	Crane	Master
Dollars per occurrence	8180	7998	7943	7925	7761	7741	7658

Table 2: Drivers Hardest to Fill Tags

Rank	1	2	3	4	5	6	7
Tag	Programming	Shift	Crane	Trainer	Associates Degree	Welding	Heavy Equipment
Time to fill	6.84	6.77	6.64	6.42	6.26	6.24	6.17

Figure 1: Drivers Most Earning Tags by State

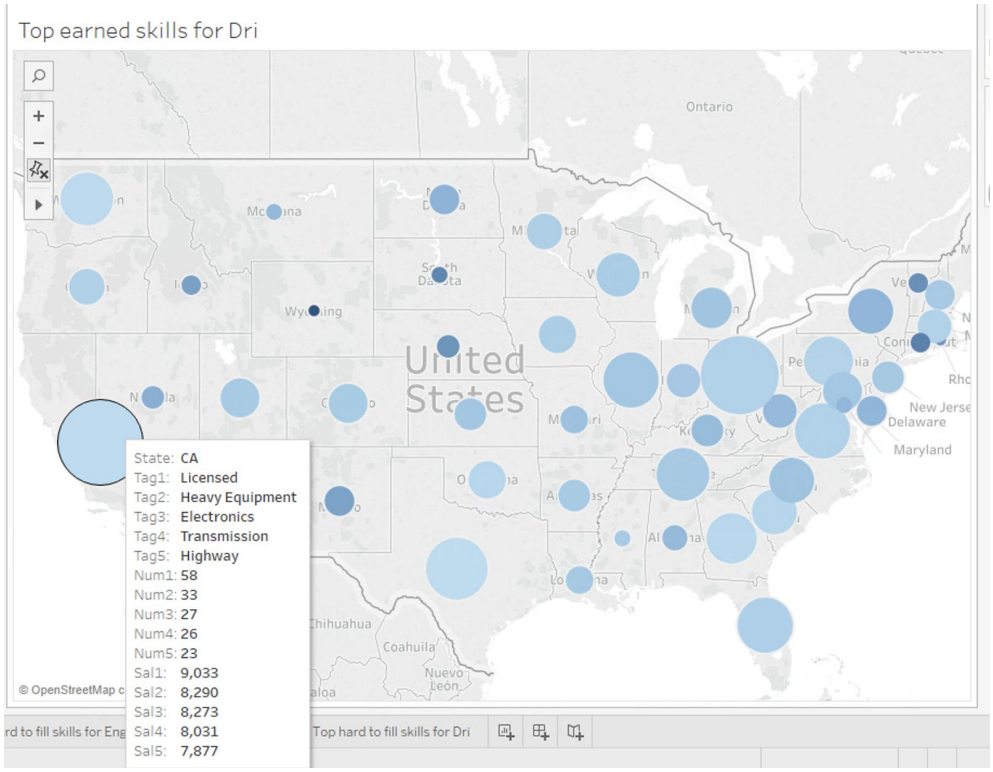
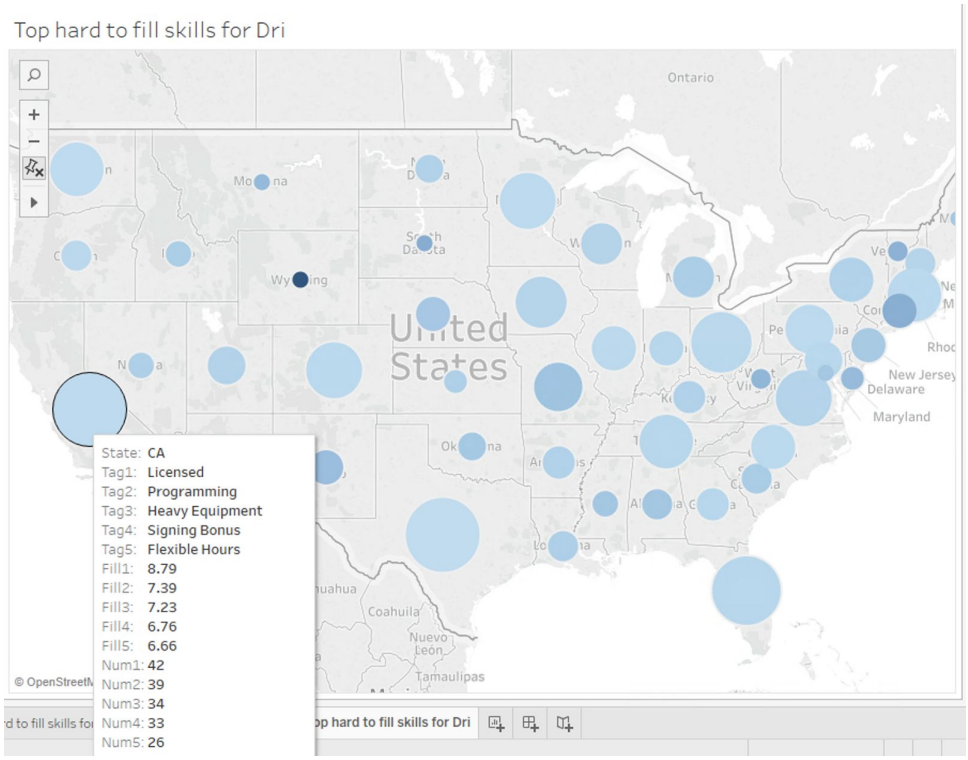


Figure 2: Drivers Hardest to Fill Tags by State



Summary 1: Multiple Linear Regression on Time to Fill of Denver Drivers Market

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.54908    0.31744  14.331  < 2e-16 ***
norm.salary                0.85111    0.27955   3.045  0.00250 **
regionAurora               0.34651    0.32822   1.056  0.29181
regionDenver              -0.41921    0.20097  -2.086  0.03770 *
safety                     0.39347    0.22224   1.770  0.07751 .
diploma                    1.05520    0.38807   2.719  0.00687 **
transportation             0.91434    0.28463   3.212  0.00144 **
cust.serv                 -0.44010    0.32722  -1.345  0.17950
training                  -1.43880    0.44824  -3.210  0.00145 **
commercial                -0.42813    0.49133  -0.871  0.38415
warehouse                 -0.34703    0.24997  -1.388  0.16593
quarter2                  -0.26098    0.35092  -0.744  0.45755
quarter3                   0.12724    0.45914   0.277  0.78185
quarter4                  -0.05228    0.40506  -0.129  0.89738
norm.salary:regionAurora   0.41351    0.36268   1.140  0.25499
norm.salary:regionDenver  -0.38218    0.20120  -1.899  0.05831 .
norm.salary:safety        -0.36193    0.19986  -1.811  0.07100 .
norm.salary:diploma        0.38828    0.21319   1.821  0.06941 .
norm.salary:cust.serv     -0.47762    0.24039  -1.987  0.04771 *
norm.salary:commercial    -0.37505    0.26504  -1.415  0.15792
norm.salary:quarter2      -0.13468    0.25506  -0.528  0.59780
norm.salary:quarter3      -0.70752    0.32832  -2.155  0.03184 *
norm.salary:quarter4      -0.44478    0.29642  -1.500  0.13438
safety:commercial         -0.65232    0.45995  -1.418  0.15700
diploma:transportation    -0.92243    0.45400  -2.032  0.04292 *
diploma:cust.serv         -0.66330    0.42369  -1.566  0.11836
diploma:quarter2          -0.93413    0.47459  -1.968  0.04981 *
diploma:quarter3           0.34858    0.59401   0.587  0.55769
diploma:quarter4          -0.32090    0.54448  -0.589  0.55599
cust.serv:training         1.19978    0.45557   2.634  0.00882 **
training:commercial        1.20626    0.50438   2.392  0.01730 *
training:quarter2          0.55742    0.48825   1.142  0.25437
training:quarter3         -0.11592    0.62043  -0.187  0.85189
training:quarter4          1.71708    0.80824   2.124  0.03432 *
commercial:quarter2        1.22261    0.54163   2.257  0.02460 *
commercial:quarter3        0.24523    0.71038   0.345  0.73015
commercial:quarter4        0.46484    0.62556   0.743  0.45793
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.725 on 354 degrees of freedom
Multiple R-squared:  0.1888,    Adjusted R-squared:  0.1063
F-statistic: 2.289 on 36 and 354 DF,  p-value: 7.131e-05
```