Machine Learning and Big Data solutions for a market-place

Wael Mohammed

Basuraj Agrawal

Teresa Morales Gomez-Luengo

Bibind Vasu

About Field Nation

- Field Nation is a B2B marketplace, which helps connect technicians to the right customers
- Most of the tasks are related to service and installation of computer hardware and software systems like POS, Networking devices, security systems etc.
- The Platform currently has more than 100,000 Technicians and generated more than 100 Million dollars in revenue every year.

How it works?

Post work orders with the click of a button

You define the scope, location, schedule, and pay. Then, post one work order at a time or thousands within minutes- templates and mass uploads make streamlining your processes a snap.

Access a marketplace of highly skilled technicians

Choose the best candidate for the job from the technicians responding to your work orders. View each technician's qualifications and reviews before reaching out or sending them on-site.

Save your trusted technicians in Talent Pools

Create pools of your favorite technicians based on skill set, location, type of work, or any other custom criteria. Instantly and reliably route work to your previously vetted and approved technicians.

Project Summary

Field Nation, a marketplace for technical services, has provided us with their 13 years of raw data, which includes:

- Work Scope: unstructured data (text) with characteristics of the work orders written by buyers
- Marketplace: extensive data about bidding and matching process, location, urgency, length, price, work outcome
- Reviews and rating

The objective of our project is to provide Field Nation with an initial impression of using alternative Big Data platforms and Machine Learning solutions for analytics.

In particular, we performed:

Exploratory Data Analysis:

- Trends over time
- Distribution of Work Orders and gross revenue by type and geography
- Graph model analysis, to understand connections between buyers and providers.
- Pricing exploratory analysis.

Business solution 1: Sentiment Analysis on customer reviews to understand the customer sentiment, post completion.

Business solution 2: Topic modeling, identifying topics from full description of the task, using LDA

Business solution 3: A pricing model to help buyers determine their initial bidding price, using topics above as input.

Data source and challenges

Data source:

- Total data spans over 40GB (13 years). Subset of ~16GB was used by us.
- All data is under an NDA and we were required to move it to our private S3 Bucket. The transfer was done by extracting data from multiple tables directly from the companies database and storing it in flat files. Upload speeds were limited to 750 Kbps!
- We setup a Databricks environment on another bucket and both buckets were linked. To reduce load time, we later saved it as tables then in HDFS (since the data was 15GB+). The files were saved as parquet files by HDFS in the same S3 bucket

Data challenges:

- *Interpretability:* Large number of attributes like type of work, description (short and large texts), pricing, location etc.
- *Format:* Most text fields contained extra delimiters, HTML tags etc, which were extremely difficult to work with. Also, since the data formats were changed recently, it was difficult to work with them without preprocessing them extremely well. At times, since the delimiters were unexpectedly present, the file reading had to be processed in chunks, while processing each line at a time.
- *Size:* The NDA in place made it more difficult, since use of RCC was not a choice. The processing of extremely large (text) data brought in new challenges, which were the most important learning bit for us in this project.

Infrastructure & Analysis Framework

Obtain ---------- Load ---------- Explore ----------→ Interpret

Data Reduction
Load to S3
Saved a copy on Hive

Amazon S3

PySpark

Data cleaning, Analysis & Modeling run using Databricks environment on AWS (PySpark, Spark 2.5)

Amazon S3

Amazon EC2

PySpark

databricks

Amazon EC2 ×8

Autoscaling Cluster (Min – 6 Nodes, Max – 10 Nodes, Mostly at 8 Nodes)

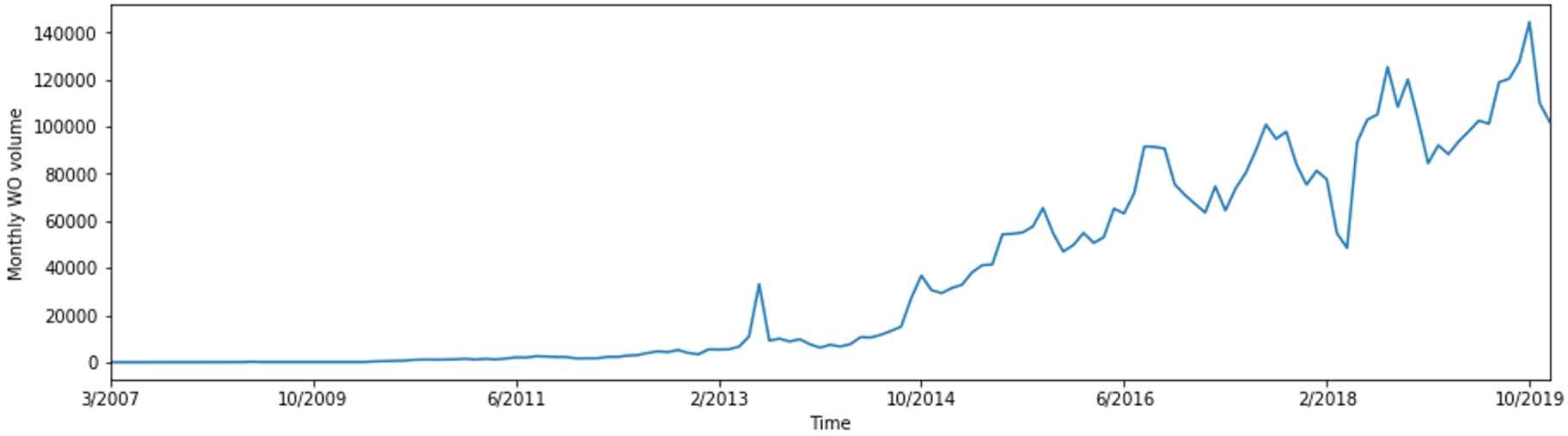
Multiple Models designed solely on Pyspark

- Sentiment Model
- Topic Model
- Pricing Model
- Hybrid Pricing Model

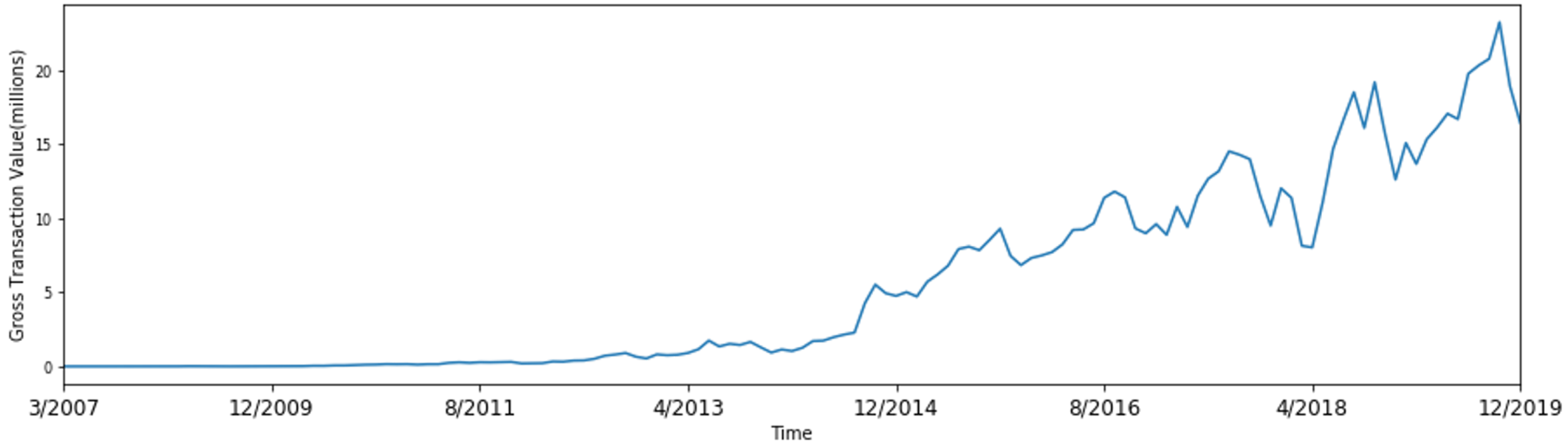
Final Interpretation using Spark & Python Graphs, as well as Graph DB on Spark

- Connected Graphs
- Python based Graphs
- Recommendations

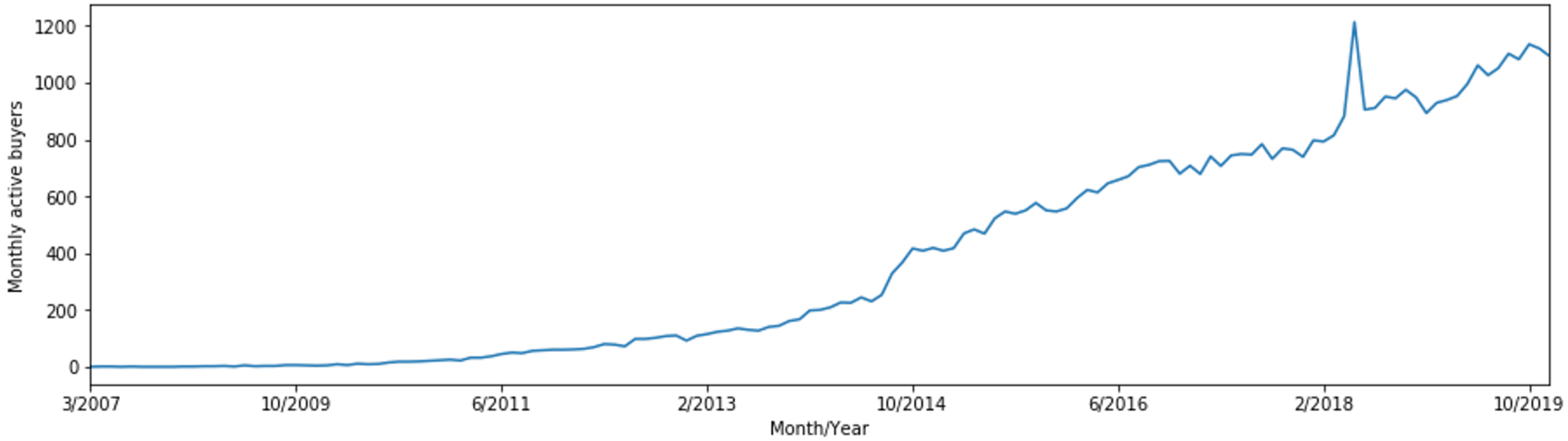
Exploratory Data Analysis and Graph modeling



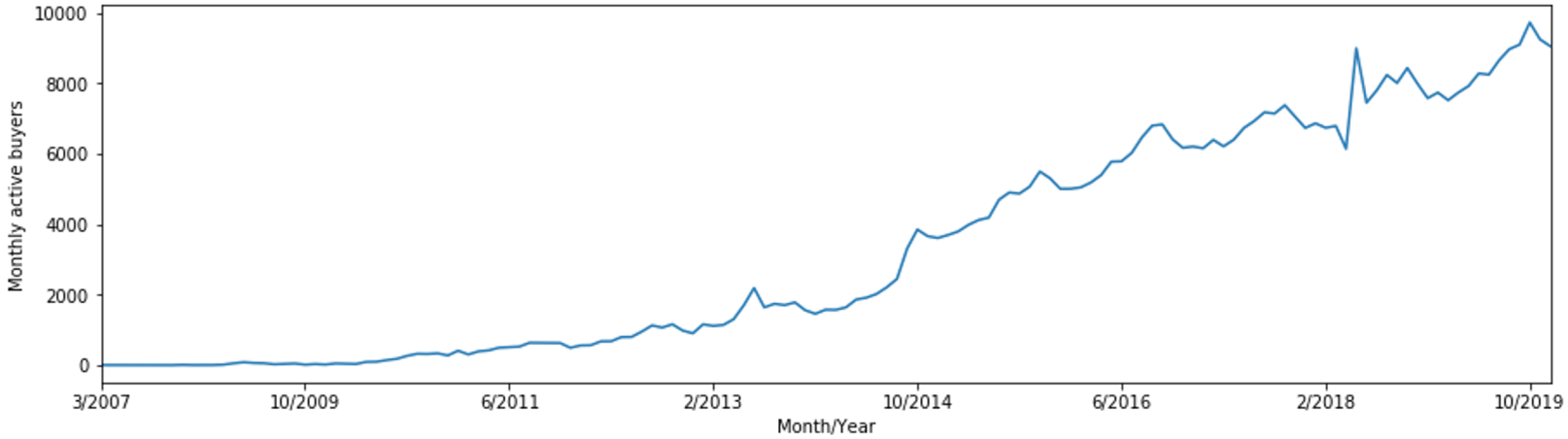
Marketplace transaction volume growth since inception



Marketplace Gross Transaction Value (GTV) growth since inception

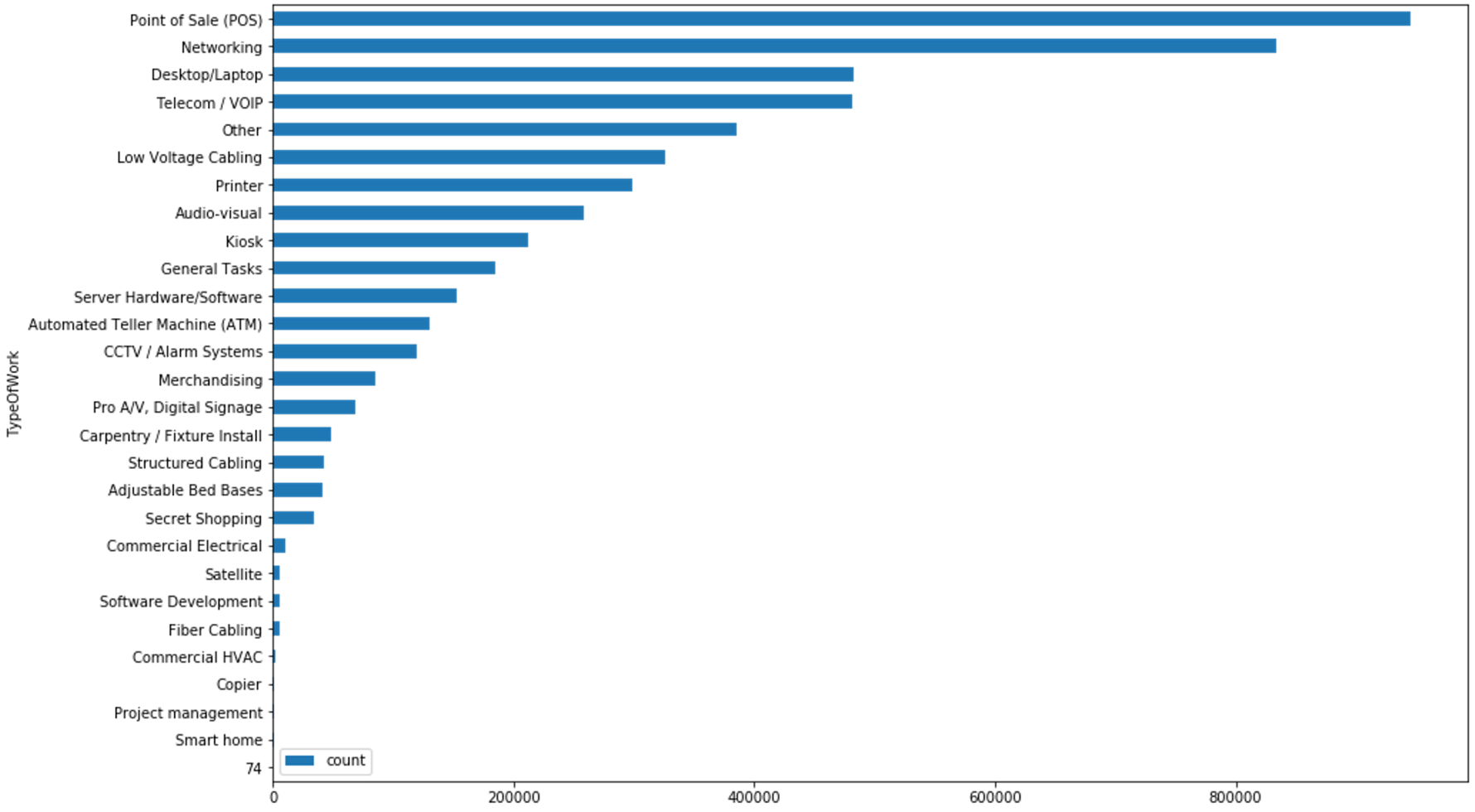


Marketplace monthly active buyerssince inception and through 2019



Marketplace monthly active providers since inception and through 2019

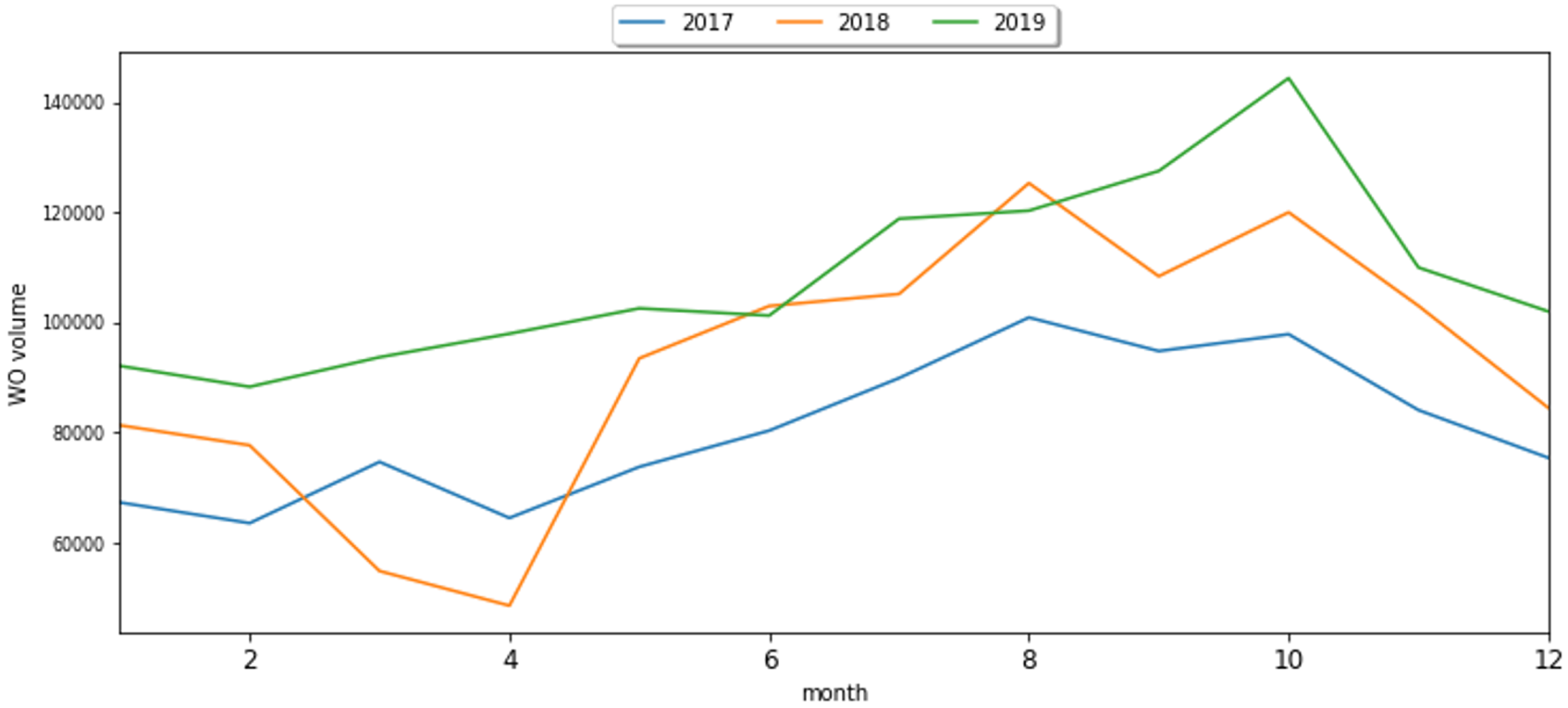
Exploratory Data Analysis - transaction volume by type of work



Marketplace transactions by types of work. Data since 2007 through 2019.

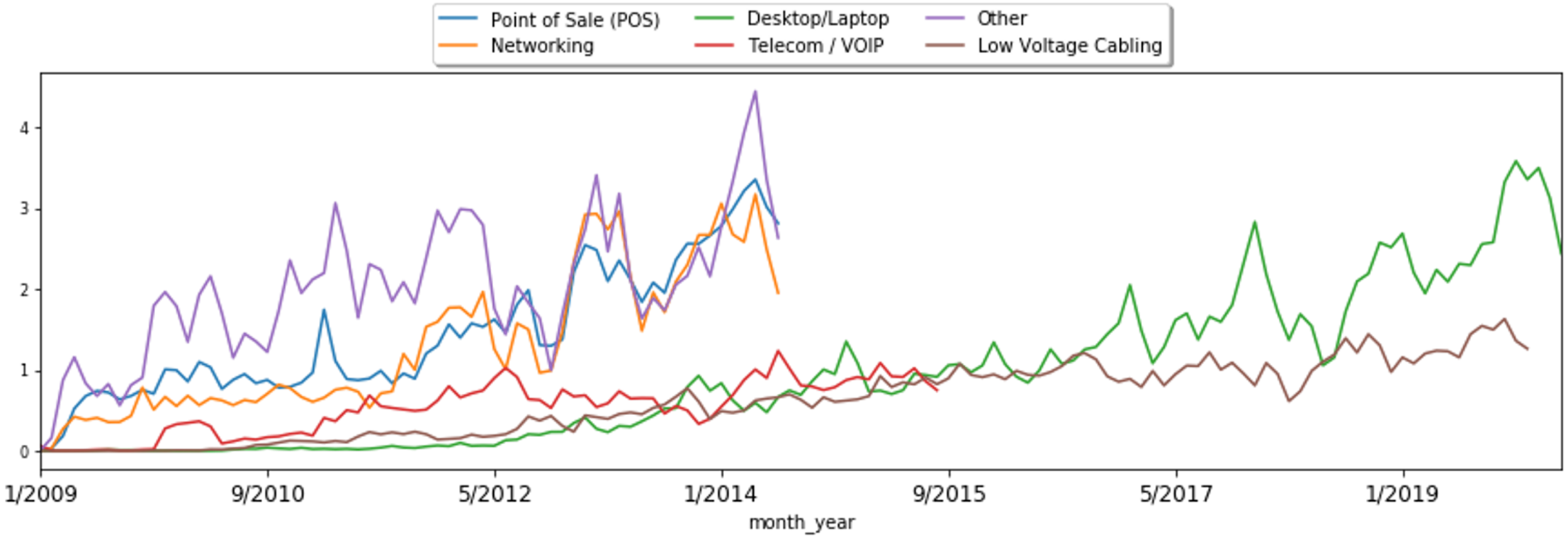
Point of Sale, Networking, Desktop/Laptop, Telecom are major types of work on marketplace.

Exploratory Data Analysis - YoY GTV growth - (2017 through 2019)



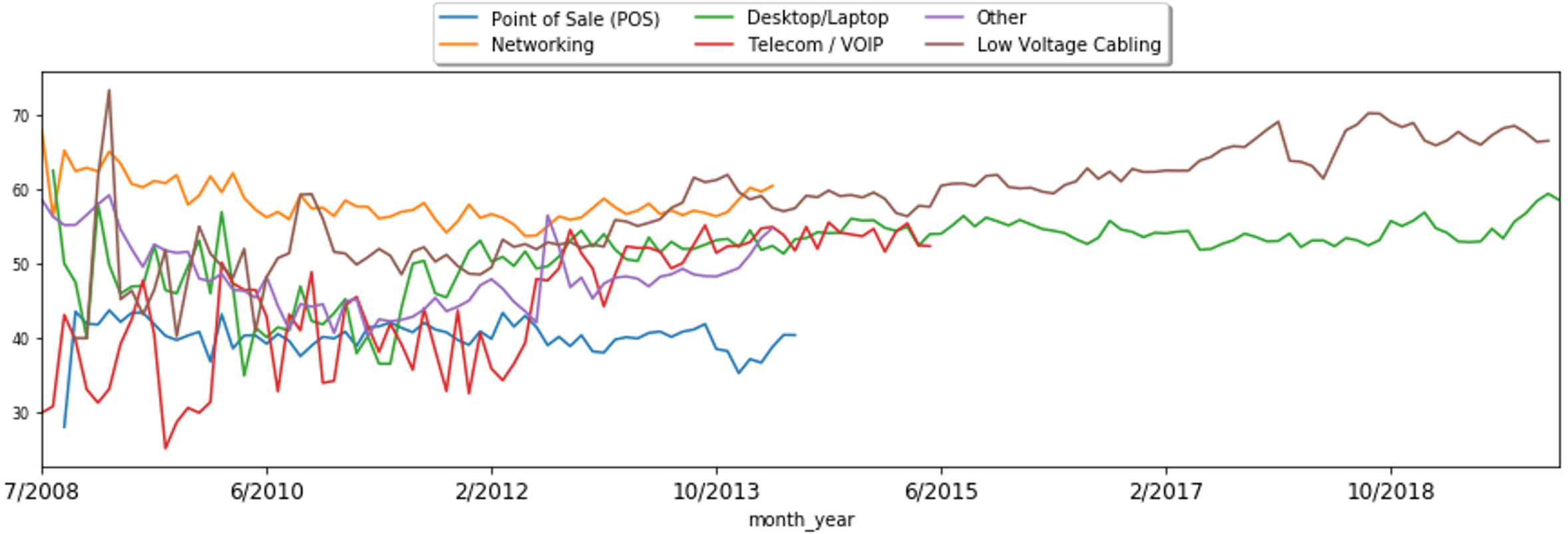
Marketplace transaction volume Year over Year growth.

Exploratory Data Analysis - GTV growth by type of work



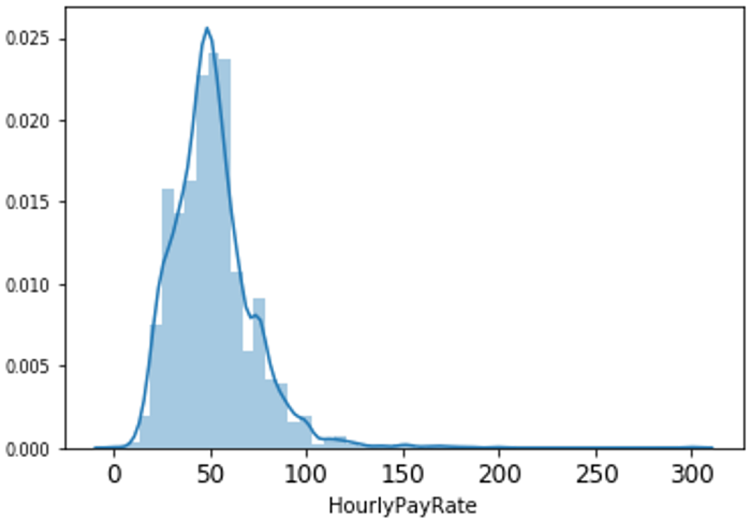
Marketplace gross transaction value growth by major types of work. Inception through 2019.

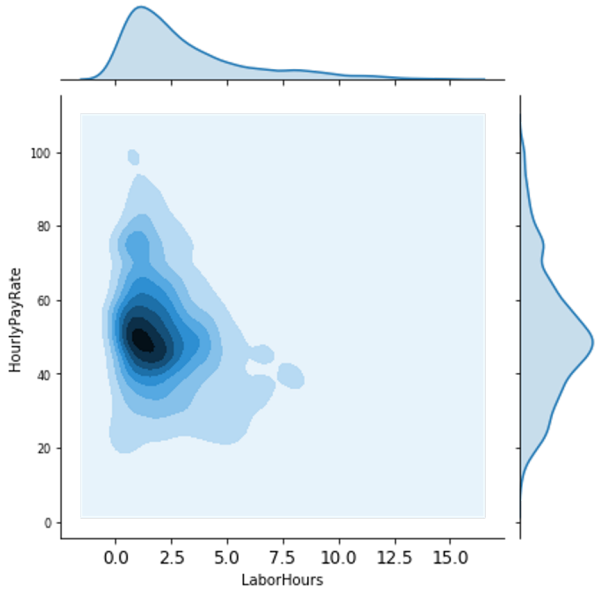
Exploratory Data Analysis - Hourly rates evolution by type of work

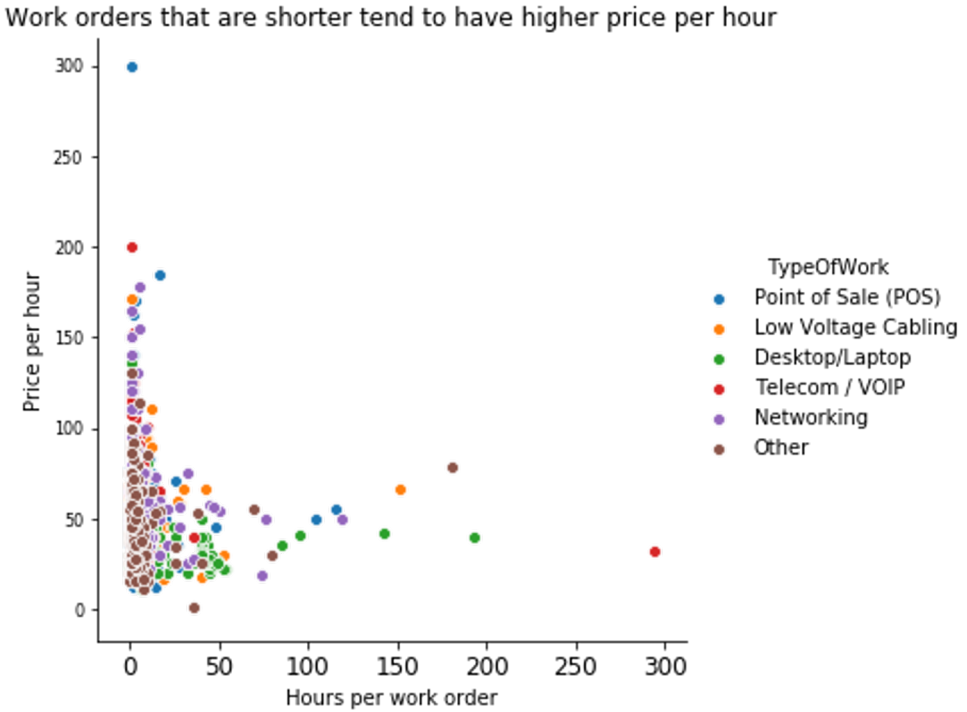


Marketplace hourly rate evolution by type of work. Inception through 2019

Exploratory Data Analysis - Hourly rates analysis by duration of work order

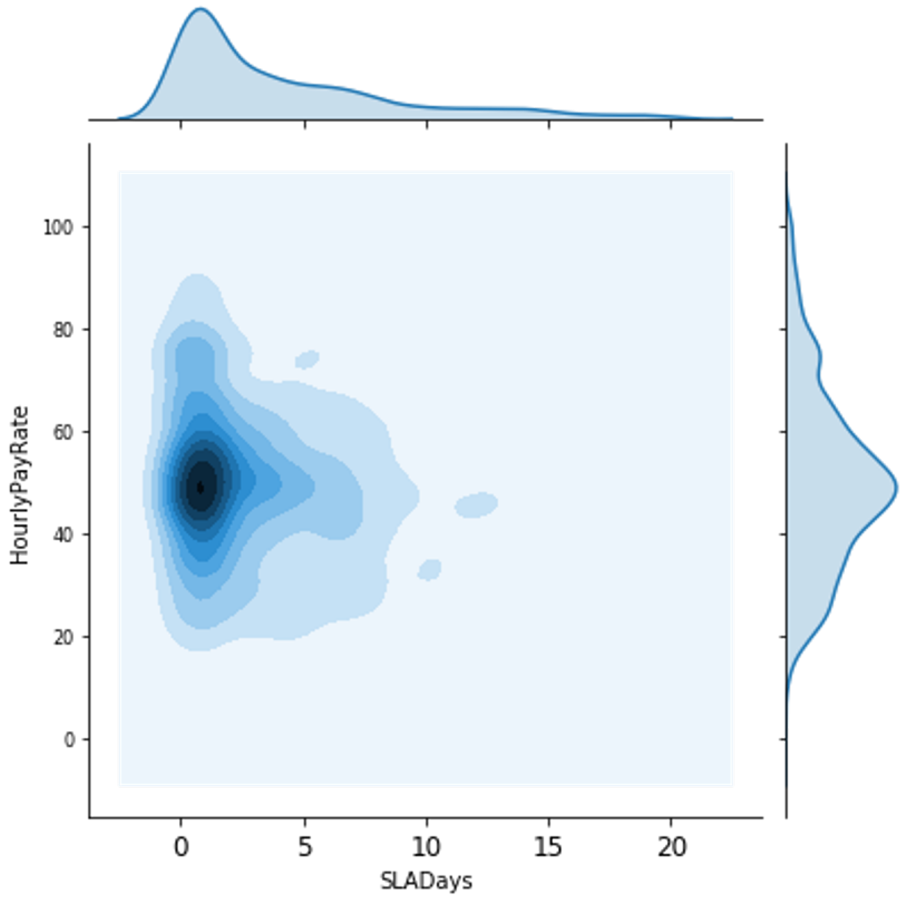


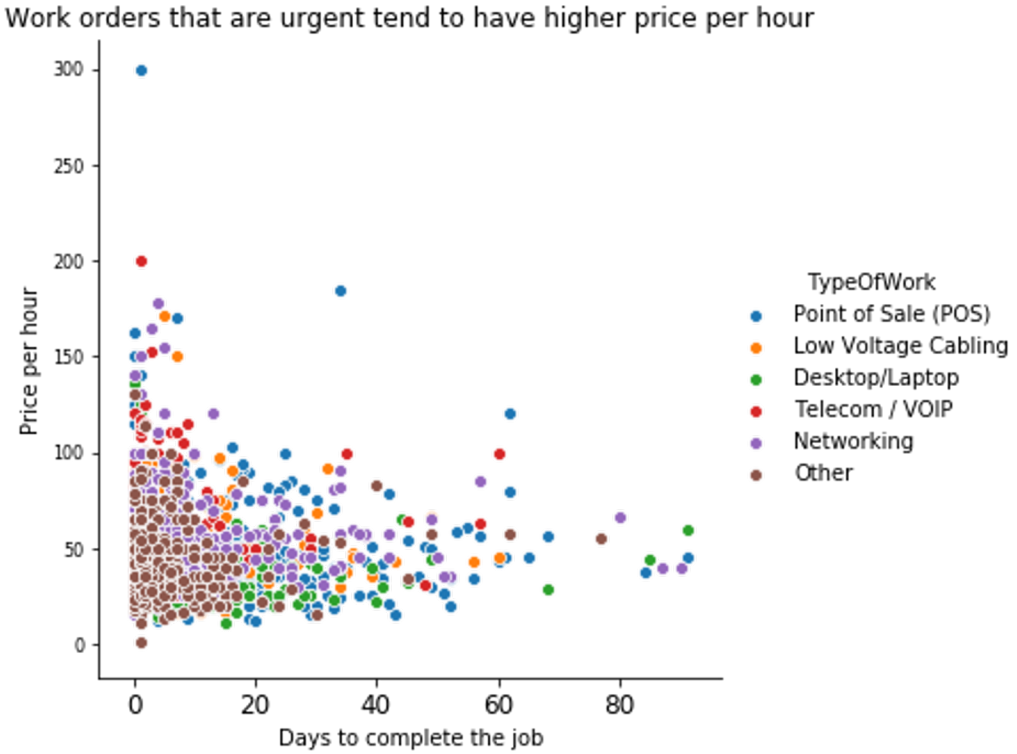




Marketplace hourly rate distribution by work duration.
Shorter duration work orders tend to have higher hourly rates.

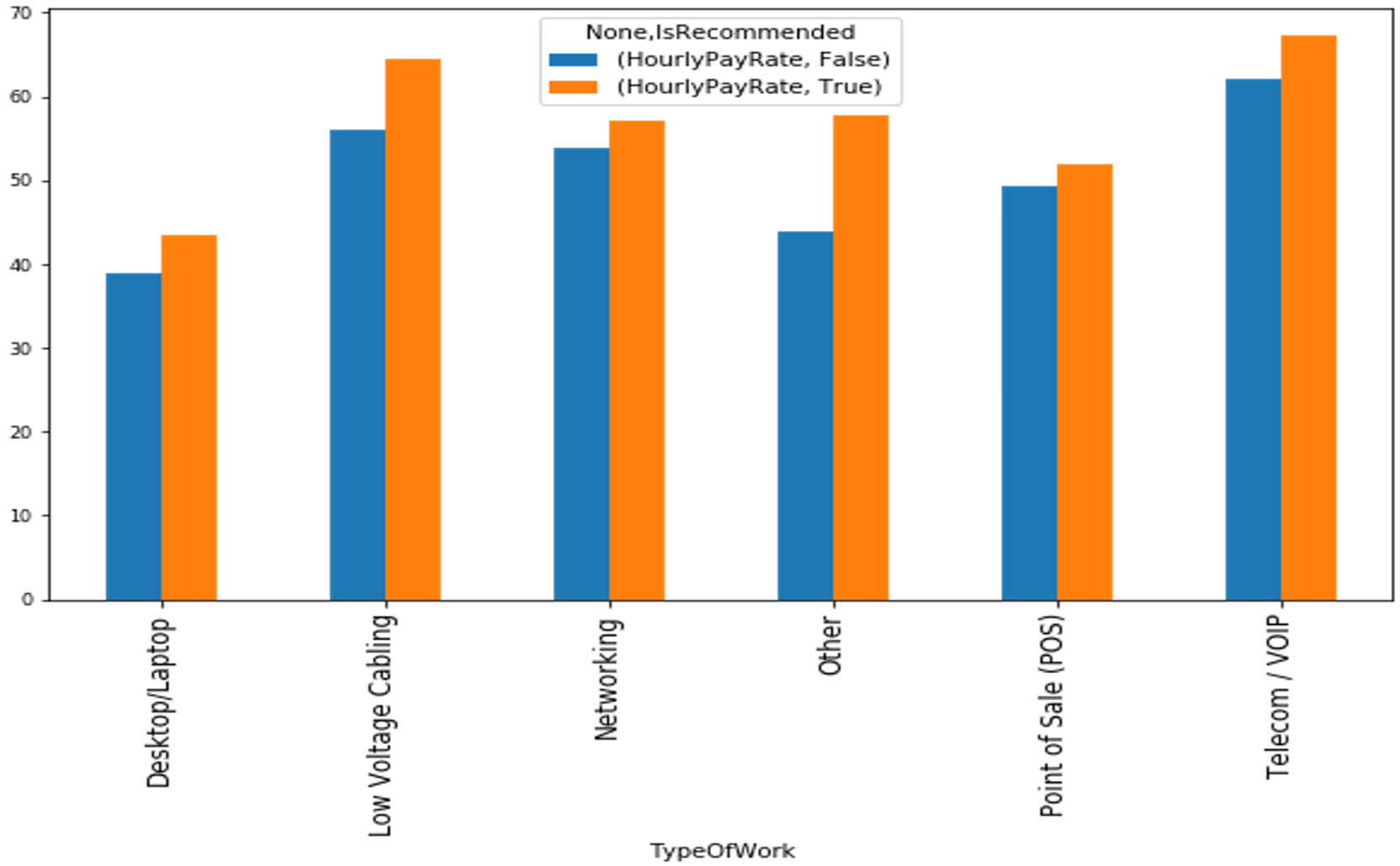
Exploratory Data Analysis - Hourly rates analysis by SLA





Marketplace hourly rate distribution by SLA.
Shorter-term (urgent) work orders tend to have higher hourly rates.

Exploratory Data Analysis - Hourly rate analysis by recommended vs other providers

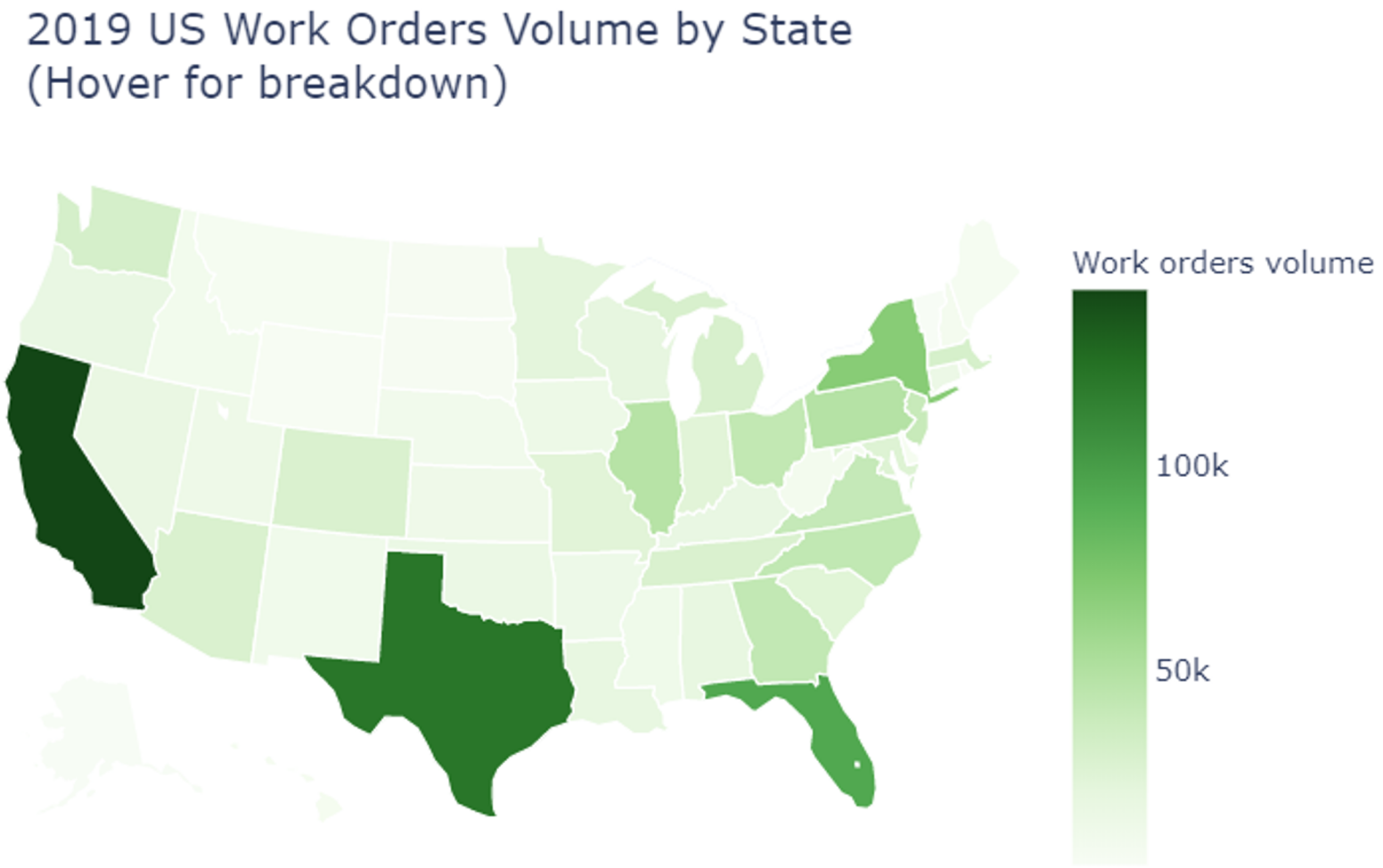


Providers recommended by the platform tend to have higher hourly pay rate than all other providers.

Data for 2018 & 2019 (since recommender system started).

2019 marketplace volume of work concentration by state. CA, TX, FL, NY, IL are high volume contributing states.

This correlates with states contribution to US GDP.

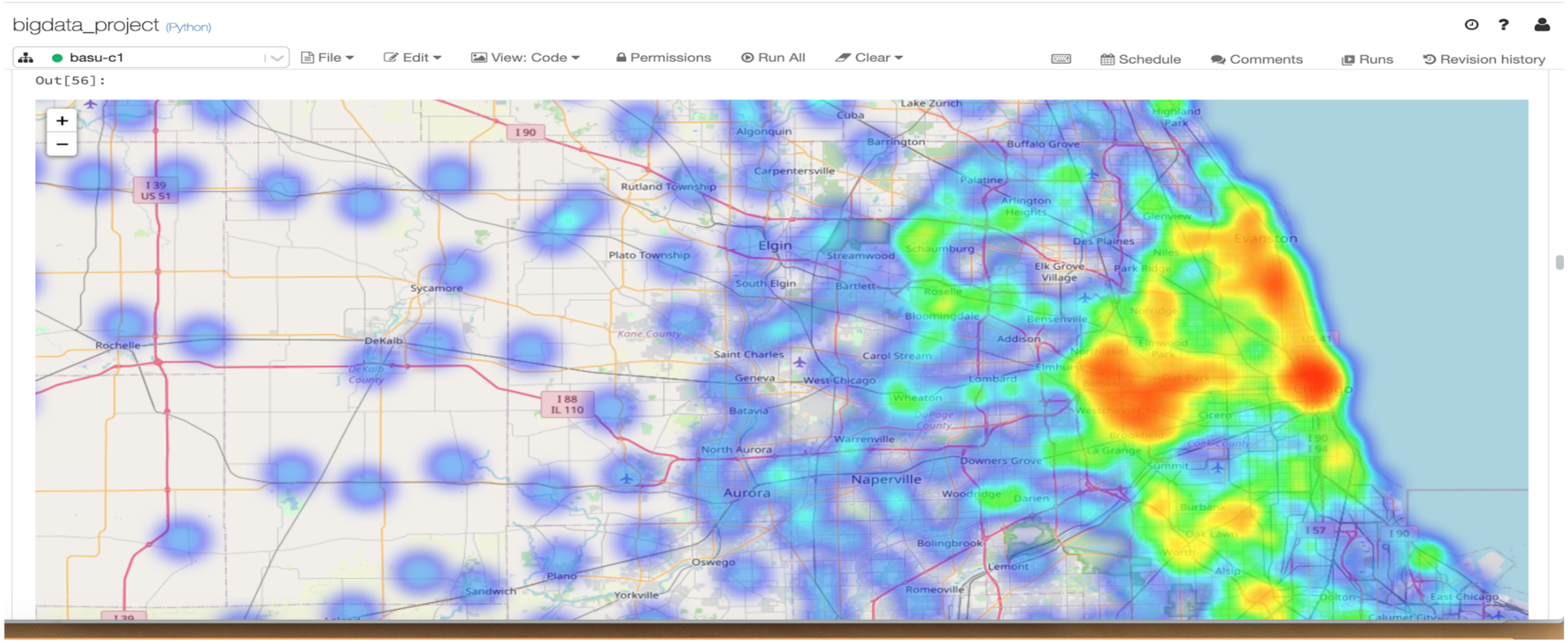


Exploratory Data Analysis - volume concentration by geo

Spatial Data Analysis of Market Place Data

1. Heat Map to visualize the Zipcodes where FN has presence
2. Attributes like Revenue, Dates visualized spatially

Use Case- Recommend areas of opportunity and strategize business development



Library: Folium, Python

- **Marketplace** (essentially a social graph) data is highly conducive to **graph model analysis.**

- Vertices (two node types):
  - Buyer (demand side)
  - Provider (supply side)

```
+-----+--------+
|   id|NodeType|
+-----+--------+
|12681|   Buyer|
| 2677|   Buyer|
|21338|   Buyer|
|  369|   Buyer|
|23679|   Buyer|
+-----+--------+
only showing top 5 rows
```

```
+-----+--------+
|   id|NodeType|
+-----+--------+
|73932|Provider|
|18119|Provider|
|37223|Provider|
|21154|Provider|
| 4752|Provider|
+-----+--------+
only showing top 5 rows
```

39876 vertices (service buyers and providers)

4352994 edges (work orders transacted since inception)

- Edges in the marketplace are relationships. **Essentially work orders!**

▸ (1) Spark Jobs

▸ edges_df: pyspark.sql.dataframe.DataFrame = [buyerid: string, providerid: integer ... 10 more fields]

buyerid	providerid	type_of_work	relationship_month	relationship_year	zip	city	state	est_revenue	distance	src	dst
188	191151	Kiosk	4	2016	36104	Montgomery	AL	140.0	9.97	188	191151
7019	35328	Point of Sale (POS)	5	2016	12804	Queensbury	NY	85.0	43.84	7019	35328
73	15880	Printer	5	2016	24504	Lynchburg	VA	135.0	6.42	73	15880
17439	90965	Networking	5	2016	94568	Dublin	CA	220.0	30.21	17439	90965
1352	65922	Point of Sale (POS)	5	2016	71203	Monroe	LA	370.0	101.17	1352	65922

only showing top 5 rows

```
1 # How many unique relationships total - since inception?
2 g.edges.select('src','dst').distinct().count()
```

▸ (1) Spark Jobs

Out[23]: 680687

680k unique relationships since inception

- Marketplace enjoys a high number of trusted relationships!

src	dst	count
57099	309670	6038
4347	76929	4747
57099	309614	3551
697	19006	3054
57099	310193	2625
255	15899	2474
193	81348	2291
57099	309869	2181
13466	124792	2139
57099	310211	2095
10157	56373	2085
697	51565	1970
57099	310055	1966
57099	315176	1950
57099	309817	1937
2667	19195	1892
13480	227168	1862
57099	407307	1770
57099	314620	1692
57099	310267	1669

only showing top 20 rows

repeat use relationships are highly healthy for the marketplace. These are considered relationships with 3+ work orders (edges)

251968 out of 680687 relationships are repeat use: (almost 37%).

- Marketplace enjoys a high number of trusted relationships!

src	dst	count
57099	309670	6038
4347	76929	4747
57099	309614	3551
697	19006	3054
57099	310193	2625
255	15899	2474
193	81348	2291
57099	309869	2181
13466	124792	2139
57099	310211	2095
10157	56373	2085
697	51565	1970
57099	310055	1966
57099	315176	1950
57099	309817	1937
2667	19195	1892
13480	227168	1862
57099	407307	1770
57099	314620	1692
57099	310267	1669

only showing top 20 rows

repeat use relationships are highly healthy for the marketplace. These are considered relationships with 3+ work orders (edges)

251968 out of 680687 relationships are repeat use: (almost 37%).

- Marketplace enjoys a high number of healthy relationships with high dollar value!
- We will measure the value of these relationships for revenue year 2019

(1) Spark Jobs
freq_rel_2019_rev_value_df: pyspark.sql.dataframe.DataFrame = [src: string, dst: integer ... 1 more fields]
freq_rel_2019_work_df: pyspark.sql.dataframe.DataFrame = [src: string, dst: integer ... 1 more fields]
valuable_rel_2019_df: pyspark.sql.dataframe.DataFrame = [Buyer: string, Provider: integer ... 2 more fields]

Buyer	Provider	2019 estimated revenue	2019 Work Orders
50143	414560	571819.504699707	105
1352	36337	500292.0	2
782	24002	465900.0	29
782	297577	461836.5	27
587	441025	408534.0	56
50143	412852	374756.6085205078	125
50143	426608	298971.3896484375	73
4936	261719	286278.1300048828	676
73200	310931	241403.5	557
2362	53285	214497.99983978271	280
4552	272859	213504.25	1027
1352	67782	195920.0	654
22077	290230	190540.64978027344	90
76549	23340	177711.0	503
56311	423717	172333.0	525
36367	14085	158504.75	123
57099	309670	145135.8800201416	611
4577	24759	144929.69999694824	328

in 2019, graph shows that marketplace had 249,848 unique relationships

Out of these relationships. 3168 had value $10k+ (1.25%)

There is a long tail of small value relationships.

Increasing high value relationships is a key strategy for the marketplace.

Key learning points:

- ➔ lkajñlkjñlkjñlkj
- ➔ lkdjalñkjñlakj

Recommendations

- ➔ lkajñlkjñlkjñlkj
- ➔ lkdjalñkjñlakj

Sentiment Analysis

1. Sentiment Analysis
    a. Regex Preprocessing
    b. Sentence Tokenizing
    c. Word Tokenizing
    d. Lemmatization
2. Word Cloud
3. Classification Models for Provider Reviews
    a. Feature - Review Text
    b. Label- Star Rating
    c. Preprocessing
    d. Count Vectorization
    e. Logistic Regression
4. Evaluation Metrics
    a. `Accuracy Score: 0.9017`
    b. `ROC-AUC: 0.8401`

Provider Reviews are Positive to a large extent

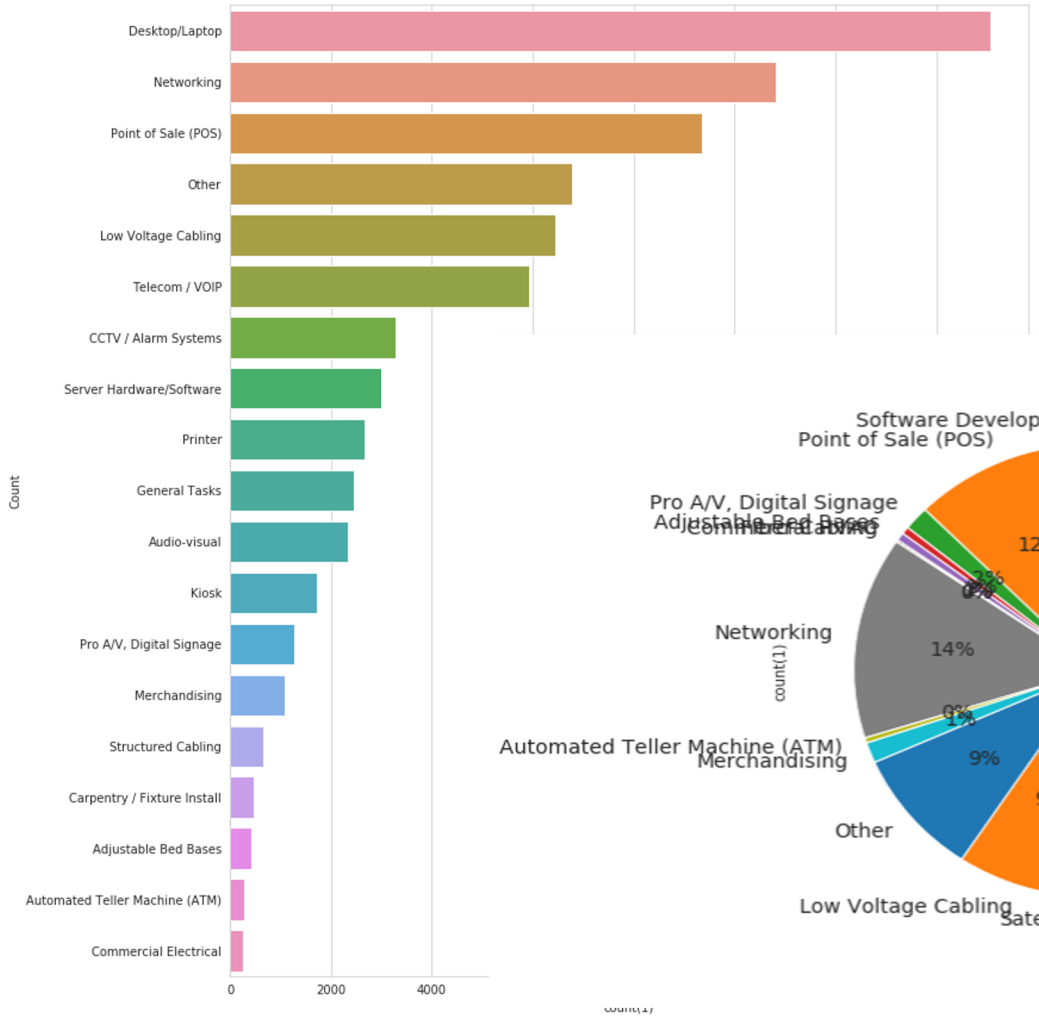
Sentiment Prediction

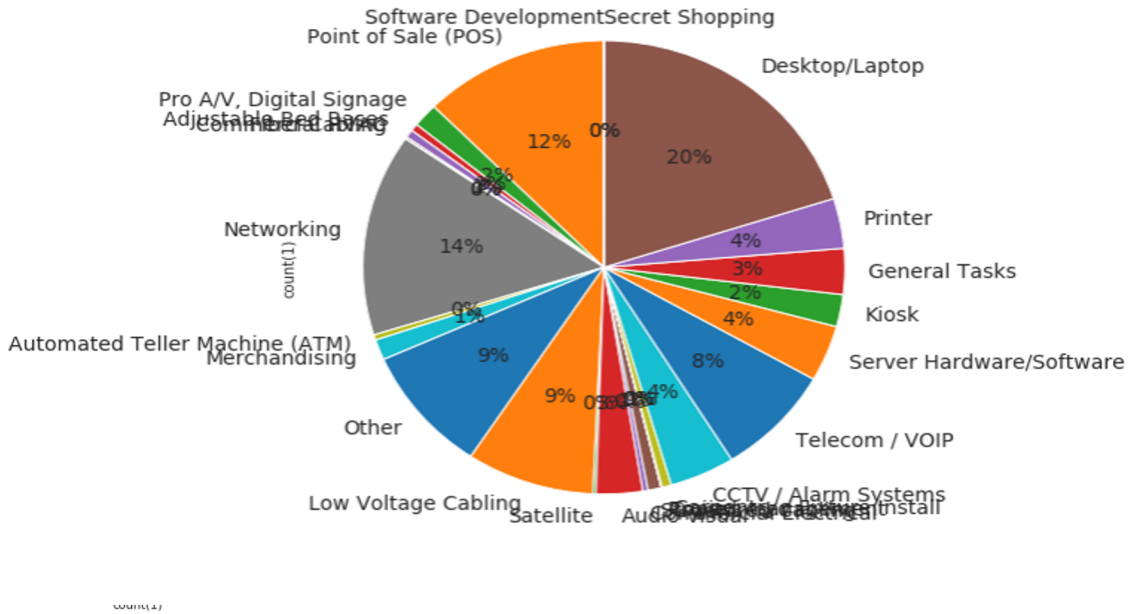
```
[[('terrific tech work', 'Positive')],
 [('comment', 'Neutral')],
 [('excellent work', 'Positive')],
 [('great job proficient', 'Positive')],
 [('darlene excellent tech', 'Positive')],
 [('rory great tech communicates', 'Positive')],
 [('comment', 'Neutral')],
 [('luis excellent job timely manner', 'Positive')],
 [('jim excellent job', 'Positive')],
 [('cws', 'Neutral')],
 [('daniel', 'Neutral')],
 [('time', 'Neutral')],
 [('use', 'Neutral')],
 [('appreciate understanding lcon situation time', 'Positive')],
 [('note client comment helpdesk manager', 'Neutral')],
 [('excellent job', 'Positive')],
 [('thanks excellent work konan', 'Positive')],
 [('great work garlando thanks', 'Positive')],
 [('excellent job tech', 'Positive')],
 [('david excellent job', 'Positive')]]
```

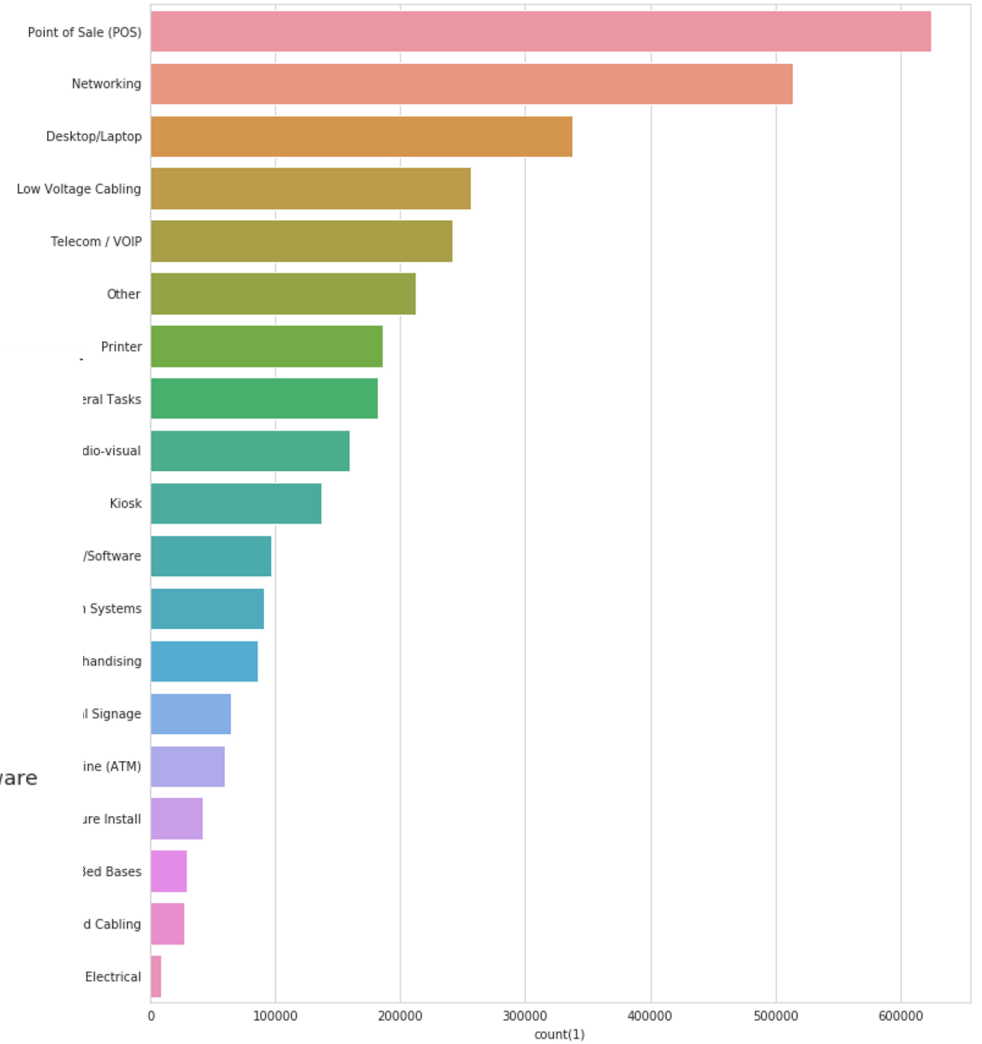
Classification of Reviews (Labels- StarRatings)

```
+----------+----------+----------+---------+-----------+--------------------+--------------------+-----+--------------------+--------------------+-----
-----+
|WorkOrderID|ProviderID|StarRating|   Review|      words|                  tf|            features|label|       rawPrediction|
probability|prediction|
+----------+----------+----------+---------+-----------+--------------------+--------------------+-----+--------------------+--------------------+-----
-----+
|    1004844|     30675|         5|  Thanks!|   [thanks!]|(65536,[38737],[1...|(65536,[38737],[3...|  0.0|[13.7671609807307...|[0.99172053995321...|
0.0|
|    1012896|     34661|         5|  Thanks!|   [thanks!]|(65536,[38737],[1...|(65536,[38737],[3...|  0.0|[13.7671609807307...|[0.99172053995321...|
0.0|
|    1013419|     16526|         5|thank you|[thank, you]|(65536,[1386,5619...|(65536,[1386,5619...|  0.0|[12.5805122435915...|[0.91555195107961...|
0.0|
+----------+----------+----------+---------+-----------+--------------------+--------------------+-----+--------------------+--------------------+-----
-----+
```

Sentiment Analysis







Provider Reviews Across “Type of Work”

Distribution of “Type of Work” in the Full Data Set

Key learning points:

- ➔ Buyers complete ratings but text reviews are rare.
- ➔ Provider reviews are positive to a large extent.

Recommendations

- ➔ Reviews should be encouraged to improve value added. One option is to incorporate “bonus points” to the recommender system when your reviews are particularly good.
- ➔ Sentiment analysis could be used to detect what types of problems are causing bad reviews and find solutions adapted to those specific types of problems

LDA topic modeling

Summary of LDA Topic Modeling Analysis –

- We use about 3 year of Scope data for our LDA analysis. To diversify and gain more insights, we run topic models on top 6 categories instead of a combined run. This contains about 3 Million records of detailed work description. It is a free text and allows users to enter up to **3000 words** per record.
- The data has a lot of pre-processing to be done before running a vectorizer. Few things we do are - remove HTML tags, web tags, symbols, line breaks, slashes, numbers, convert to string, lower casing, stopword removal etc.
- We then run a Tokenizer over it. This is then followed by a TF-IDF and an LDA process. We save the model, process the LDA across all the data and save the output in a hive table. We also convert the probabilities tables from a list using a UDF on Argmax function from Numpy.
- We make 6 different topic models, each with 7 topics. One of the categories, “Telecom & VOIP” have been described in the next page.

Topic Modeling Example - Preprocessing

HTML Tags need to be removed with the text inside it

Web tags are very difficult to clean when with symbols

Example of a cleaning segment -

Pre processing Data -

```
<p><strong>Overview of steps</strong></p><p><br></p><p><u>At the direction of the Insight Team Lead, move the old computers, monitors and printers to a storage area in the hospital. Ensure patch cords, power cables and old surge protector are included. Ensure that they are stored in a neat and organized manner.</u></p><p><strong>Thin Client Configuration</strong></p><p><u>Step 1.    Unbox the Thin Client and Monitor</u></p><p><u>Step 2.    Hook up the keyboard, mouse, power cable, USB cable, Ethernet cable, display port cable and stand (this will require a #2 Philips screwdriver) to the Thin Client. </u></p><p><u> Do not use the blue VGA cable!</u></p><p><u>Step 3.    Unbox a new surge protector and plug the thin client power brick and monitor power cord into it. </u></p><p><u> Do not use any of the hospital□s surge protectors!</u></p><p><br></p><p><strong>Laser Printer Configuration</strong></p><p><u>Step 1.    Unbox all of the laser printer and remove the following</u></p><p><u>a.    Orange tape</u></p><p><u>b.    Cardboard </u></p><p><u>c.    Sealing tape</u></p><p><u>Step 2.    Plug the printer into a new surge protector of a nearby Thin Client</u></p><p><u>Step 3.    Connect the printer to the Ethernet drop then turn on the printer</u></p><p><br></p><p><strong>Zebra Label Printer Setup</strong></p><p><u>Step 1.    Unbox the printer and load the labels</u></p><p><u>Step 2.    Plug in the power and Ethernet cables</u></p><p><br></p><p><u>The status light will change from amber to green □ wait 30 seconds to allow the printer to get an IP address</u></p><p><br></p><ul><li><u>Step 3.    Hold down the green button until it blinks once and then let go. The printer will print out a Printer Configuration and Network Configuration</u></li></ul>
```

Post processing Data (Spaces are fine, since data will be tokenized) -

```
 Overview of steps       At the direction of the Insight Team Lead move the old computers monitors and printers to a storage area in the hospital Ensure patch cords power cables and old surge protector are included Ensure that they are stored in a neat and organized manner    Thin Client Configuration    Step     Unbox the Thin Client and Monitor    Step     Hook up the keyboard mouse power cable USB cable Ethernet cable display port cable and stand this will require a  Philips screwdriver to the Thin Client      Do not use the blue VGA cable    Step     Unbox a new surge protector and plug the thin client power brick and monitor power cord into it      Do not use any of the hospitals surge protectors       Laser Printer Configuration    Step     Unbox all of the laser printer and remove the following    a    Orange tape    b    Cardboard    c    Sealing tape    Step     Plug the printer into a new surge protector of a nearby Thin Client    Step     Connect the printer to the Ethernet drop then turn on the printer       Zebra Label Printer Setup    Step     Unbox the printer and load the labels    Step     Plug in the power and Ethernet cables       The status light will change from amber to green  wait  seconds to allow the printer to get an IP address       Step     Hold down the green button until it blinks once and then let go The printer will print out a Printer Configuration and Network Configuration
```

Command took 27.80 seconds -- by Basuraj@uchicago.edu at 6/4/2020, 2:47:34 PM on Basu-C5-Mega

Spaces are okay, since we are going to Tokenize this data. Stopwords are removed while tokenizing too.

Few users write the entire text in CAPS. We convert this to lowercase while tokenizing

Topic Modeling Example - Topics (Telecom & VOIP)

- The topics for the category are shows here. We have 7 topics per type of work.
- There are well known mentions like Cisco, Bullseye, Spectrotel, BEC, CPT (code) etc along with the kind of keywords we mostly find with them. For example, Spectrotel is a reseller which sells BEC products, and seems like most of these requests are to fix BEC firewalls sold by spectrotel.
- Such a model can be used to define subcategories or suggest a pricing based on commonly used keywords together. Since we are just proposing a price range based on Keywords, the company is free from most legal requirements.

topic	words
7	[phones, lines, line, phone, end, user, install]
6	[plugs, cabinet, spectrotel, bec, cable, firewall, labeled]
5	[unit, service, return, order, installer, provider, user]
4	[tech, check, rack, must, panel, technician, cisco]
3	[circuit, rj, port, laptop, router, install, serial]
2	[bullseye, foot, per, gt, extension, expense, lt]
1	[cpt, request, job, order, support, tech, site]

Model has a low perplexity value, and they were constant across both test and train dataset. This showcases the stability as well as the generalizability of the model.

```
1 lpt, lp = ldaModel.logPerplexity(df_testing), ldaModel.logPerplexity(df_training)
2 print("Perplexity on testing and training data: " + str(lp) + ',' + str(lpt))
```

▸ (4) Spark Jobs

```
Perplexity on testing and training data: 7.332533244081446,7.343086267940751
```

Command took 17.57 minutes -- by Basuraj@uchicago.edu at 5/30/2020, 10:44:30 PM on Basu-C5-Mega

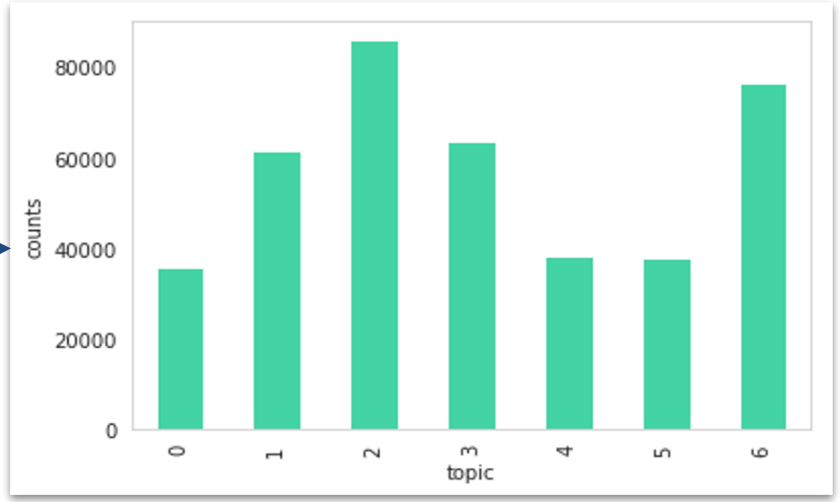
Topic Modeling Example - Single Document

The Topic Selected by the model for this item is Topic 5 (Prob - 96%). The Keywords for topic were - unit, service, return, order, installer, provider, user.

This topic seems to be used by retail partners to go ahead and “outsource” the work of finding a technician to setup new parts received. If any faulty parts are present, they are requested to mark a return as seen.

```
[(1, 0.0), (2, 0.0), (3, 0.0), (4, 0.0), (5, 0.966), (6, 0.032), (7, 0.0)]
Out[100]: [Row(descriptions='     SOW      Kindly dispatch a tech for L troubleshooting     Check the cabling between port  of SW and AP     Re terminate the cable on both ends and also recrimp the cab
le    on patch panel end   check if it works      Check if the port has any issues      Carry EG Laptop  internet extra cables cable tester  console    cable
Technician Requirements     TO RECEIVE PAYMENT FOR THIS JOB REQUIREMENTS BELOW MUST BE MET     Service call from Gap GTI Network Operations amp Support     Tech must work with GAP GTI Telephone
Required tools            Laptop with    wireless capabilities            Console cable            Basic hand tools          AFTER HOUR Emergencies affect stores ability to function    the nextsame da
y No power no registers up and working it    is MANDATORY tech speaks to a LIVE person at Spencer Technologies    before leaving the store      Tech Must check inout with Spencer service    option       T
ech Must check inout with GAP GTI       Technician must take before and after site photos of work if    scope involves a cable run or any devices in the Network Cabinet   Technician must send these phot
os to servicespencertechcom      Technician MUST write RMA number on anyall boxes being   returned to Spencer technologies Number MUST be visible and    legible   write in Marker on outside of boxes
In the event you do not have a RMA number please write the    Call number on the boxes being returned     Support      While onsite support is provided by Spencer    Support   option       If support canno
t be contacted contact Spencer    Service   option                    ')]
```

The Graph showcases the distribution of topics in various categories. As seen, the distribution is very uniform, with no topic under 35K mark. This showcases that the topics are very diverse and the model is able to cover the spread. Splitting into more counts is a possibility, but it is better done using a semi supervised topic model like CorEx to overcome spareness of that data, which LDA forms.



Key learning points:

- ➔ Most tasks are concentrated to few manufacturers and similar kind of jobs.
- ➔ Big Data is a great use-case when it comes to cleaning and doing all the pre-processing steps. It is probably faster than other technology available in the market, as well as falls cheaper than running it on GPUs. The model however can be better trained on a GPU service for 2 reasons -
  - ◆ The community support for new NLP techniques on GPU based machines are higher
  - ◆ The processing speed is faster on GPU over CPU, even though we are engaging more cores in spark. Network delay adds up to the process. However, we should definitely try spark with GPU.

Recommendations

- ➔ Pricing can be dynamic and predicted via the topic categories. This is a great use case to explore.
- ➔ A better training model could be obtained if we can spend more time collecting more industry specific keywords to remove.

Pricing model

Business Problem:

To help buyers determine what the initial price should be

Data selection and preparation:

- Subset: 6 most common types of work/3 last years
- Prediction target: Ex-post hourly price (`HourlyPayRate´)
- Features (only if available before bidding)

$$P_s = f(Q_s, P_a, C)$$

Market Supply	Alternative Price	Costs
Preferred provider	Price State (5 prev. days)	Type of Work
Requests ratio in State	Price City (5 prev. days)	Labor hours
Requests ratio in City	Dummy for year	Pay basis
Declines ratio in State	Dummy for month	Urgency (SLA Days)
Declines ratio in City	Dummy for State	Scope (from LDA)

Training:

- 70/30 training test split (460,000/ 197,000 observations)
- Tried with different regularization parameters, depth of trees and number of iterations/trees and selected the ones that did not show overfitting.

Results and model selection:

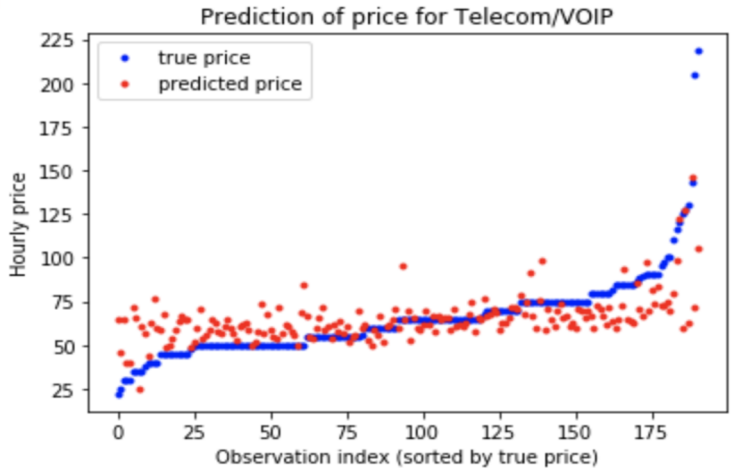
MODEL	WITHOUT SCOPE		WITH SCOPE		TRAINING TIME
	RMSE TEST	RMSE TRAIN	RMSE TEST	RMSE TRAIN	
Linear Regression (elasticnet regularization, 0.1, 0.5)	16.75	16.9	16.23	16.23	10 min
Gradient Boosting Regression (depth= 10, iter=150)	14.03	13.93	13.12	12.99	55 min
Random Forest Regression (depth = 20, trees = 100)	**13.43**	13.12	**12.52**	12.32	44 min

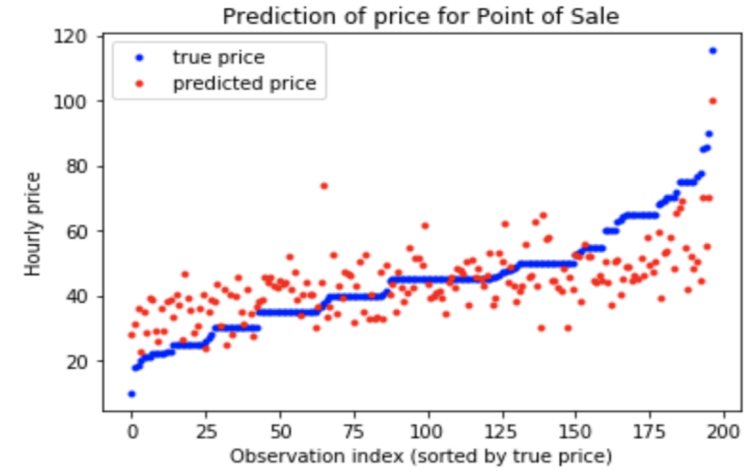
Results by type of work (best model):

Type of work	Average Price	RMSE (without scope)	RMSE (with scope)	% Change	RMSE/average price
Point of Sale	44.15	12.39	11.11	-10.3%	**25.12%**
Networking	51.73	13.78	13.38	-2.9%	25.86%
Low Voltage Cabling	54.18	14.76	14.48	-1.9%	26.68%
Telecom/VOIP	64.40	16.38	14.60	-10.9%	**22.78%**
Printer	50.07	13.77	13.43	-2.5%	26.64%
Desktop/Laptop	37.13	10.08	9.60	-4.8%	25.89%

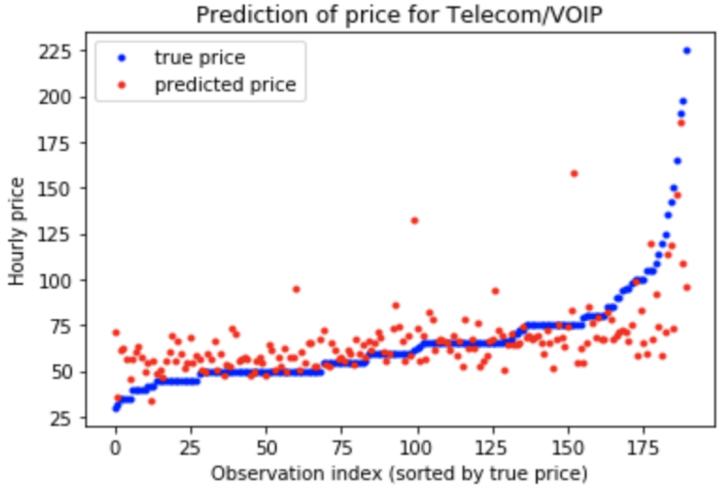
Results by type of work (cont):

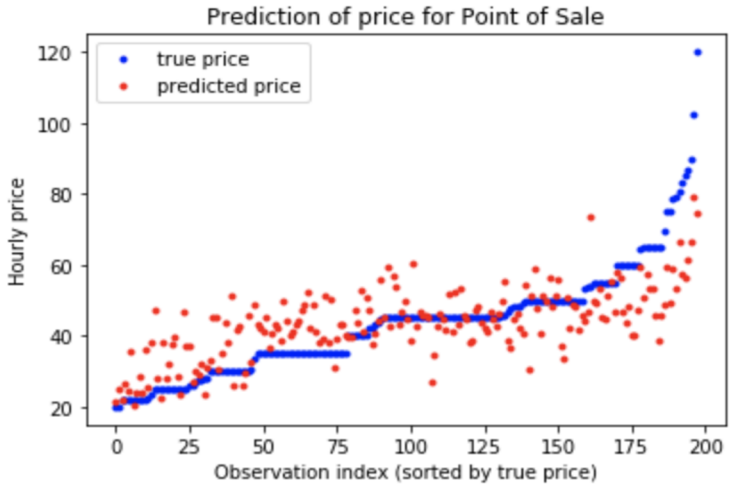
Model without Scope





Model with





Key learning points:

- ➔ The scope can be used to understand the nature and cost structure of WOs.
- ➔ Estimated prices seem to be within a reasonable range, particularly for telecom and POS.
- ➔ Model can probably be improved with:
  - ◆ Improved LDA analysis of the scope
  - ◆ Adding more data such hour in which the job is posted, other estimates of market price and costs;
  - ◆ Performing analysis of outliers by type of work and State.

Recommendations

- ➔ Data collection should be adapted for implementation and improvement of the model. This includes asking buyers to add estimated hourly price for any paybasis and estimated hours.

- ➔ Pilot program with one of the work types could help to understand the impact of this model on:
  - ◆ Customer satisfaction.
  - ◆ Field Nation’s Revenue

Future Scope

Scope of Improvement

LDA Model:

- Using semi—supervised topic model like CorEx to predefine a few categories, while exploring new keywords.
- Refine topics by pending more time on cleaning industry specific keywords, after multiple iterations.
- Expand to all types of work.

Pricing Model:

- Refine features, impute missing values and add new features such as hour, characteristics of buyer/provider...
- Neural networks
- Expand to all types of work

Appendix

Script names

1. **Exploratory Data Analysis:** “Data_Exploration”

1. **Graph model:** “Graph_Ex1”

1. **Sentiment analysis:** “bigdata_project”

1. **LDA model:**
    a. “Lda_Pyspark_networking_example_final”
    b. “lda_pyspark_automated_final”

1. **Pricing model:**
    a. “Pricing_model_final”
    b. “Pricing_results_visualization”
    c. “Pricing_results_visualization_LDA”