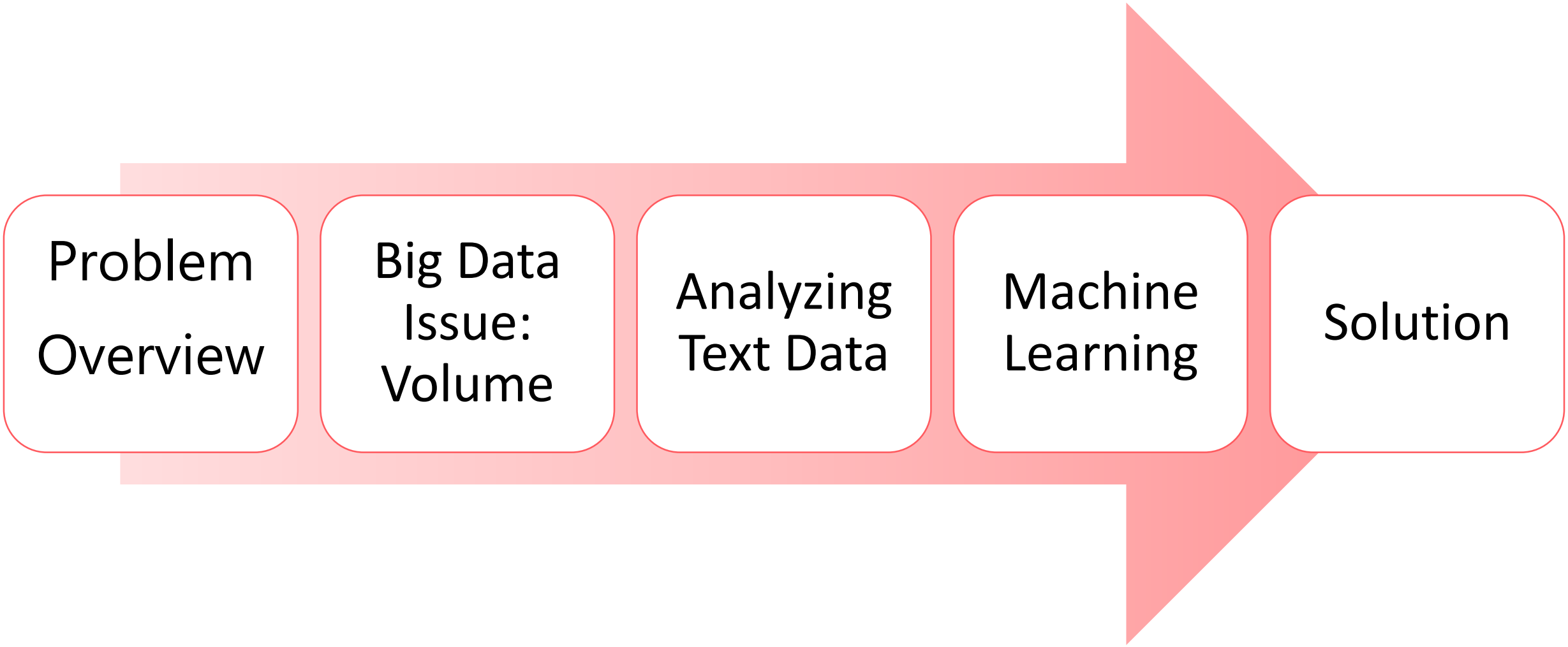
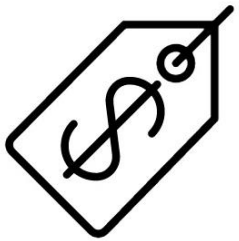
Breaking Even With

[redacted] & Victoria Sharam

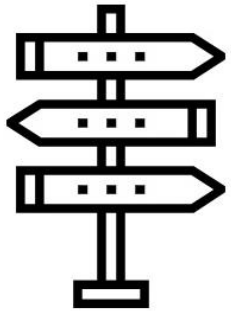

Problem Overview
Big Data Issue: Volume
Analyzing Text Data
Machine Learning
Solution

How can an NYC Airbnb host optimize their experience and aim to break even on average monthly rent

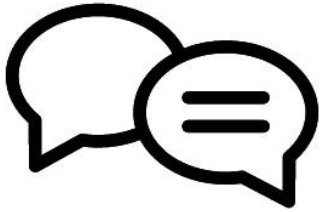
Price



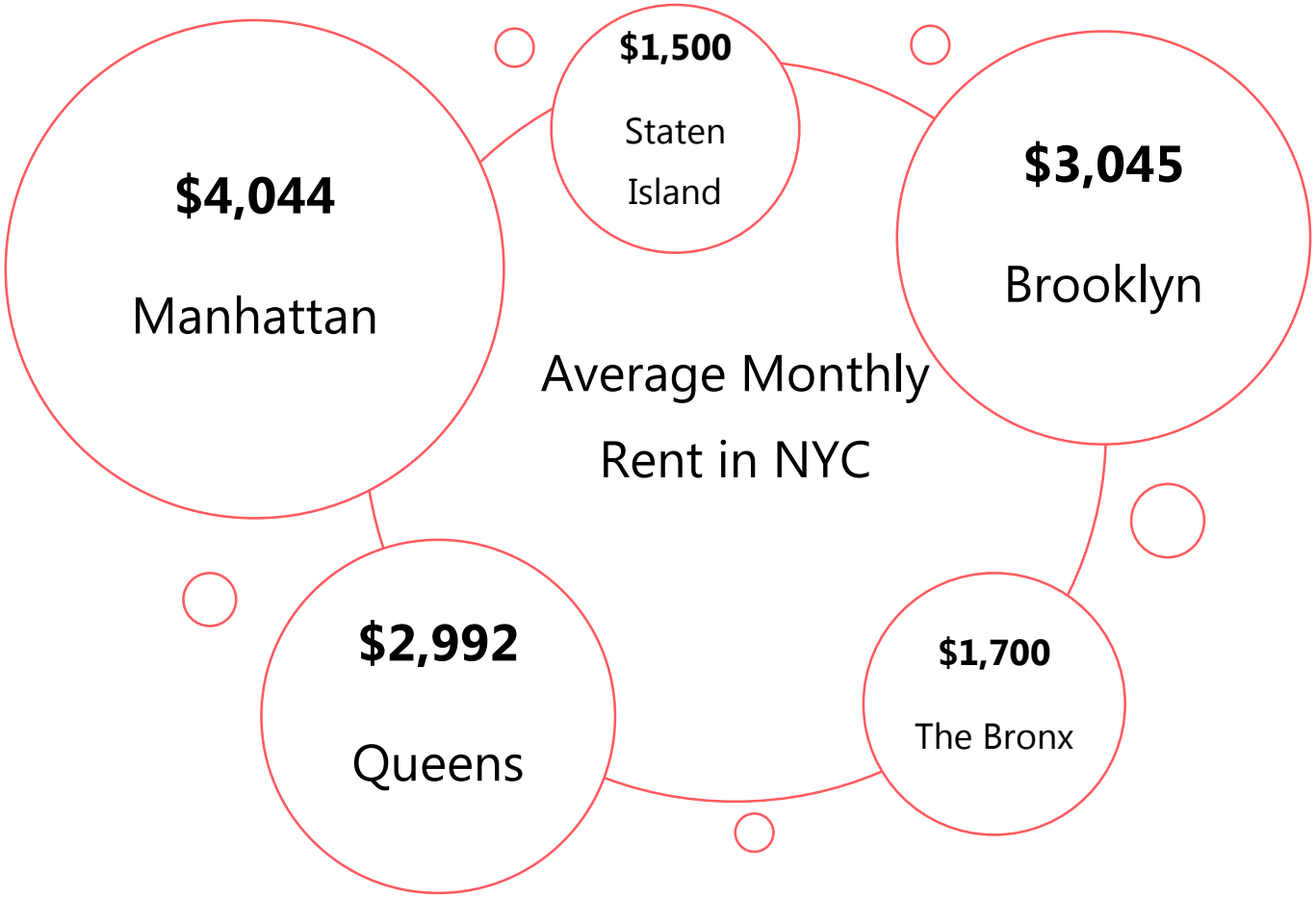
Borough



Number of Reviews



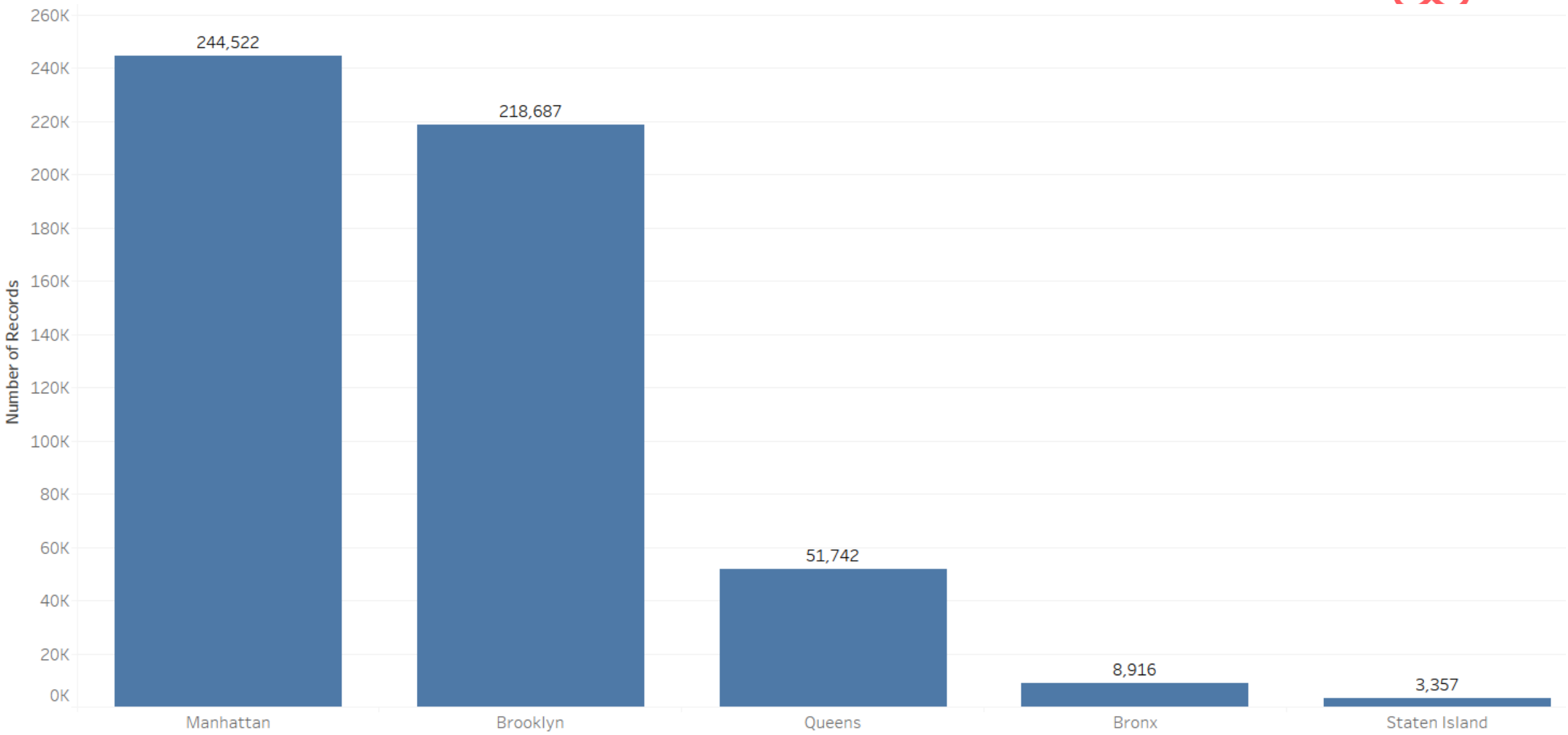
Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution



Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

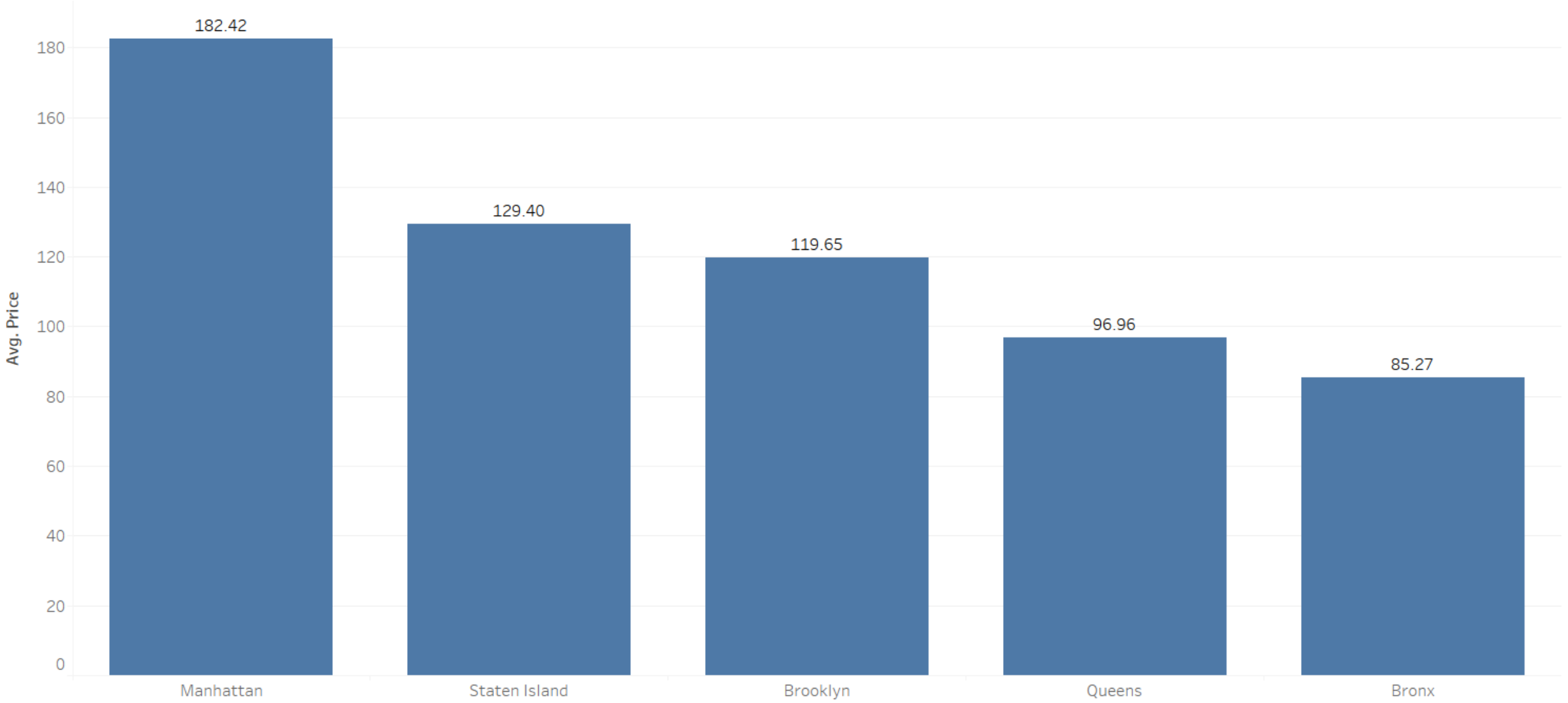
Source: https://www.brickunderground.com/rent/average-rents-nyc

Number of Listings by Borough



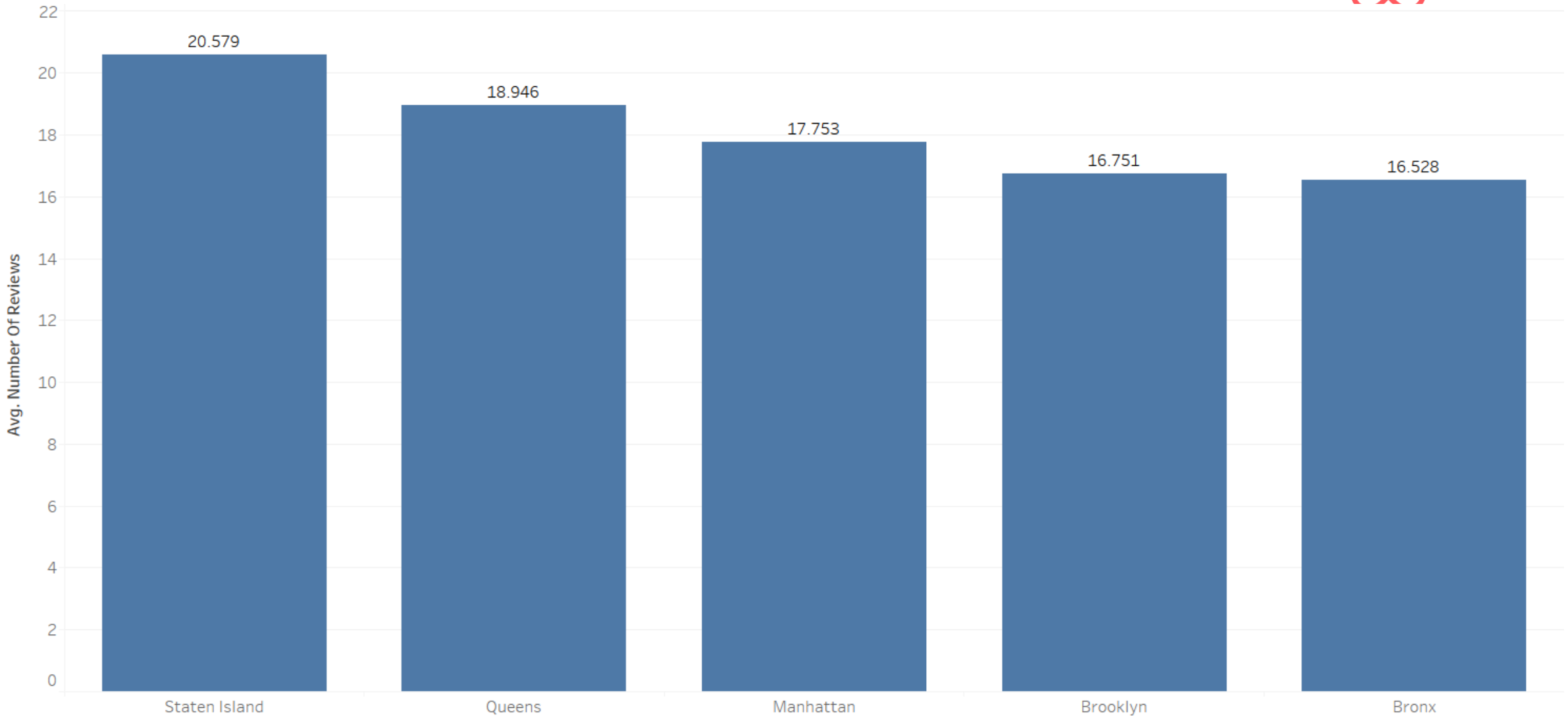
Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Average Daily Price by Borough



Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Average Number of Reviews by Borough



Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Dataset Structure

	V1	V2	V3	V4	V5	...	n variables
3/17							
4/17							
5/17							
6/17							
7/17							
8/17							
9/17							
10/17							
11/17							
12/17							
01/18							
02/19							

Principal Component Analysis

1

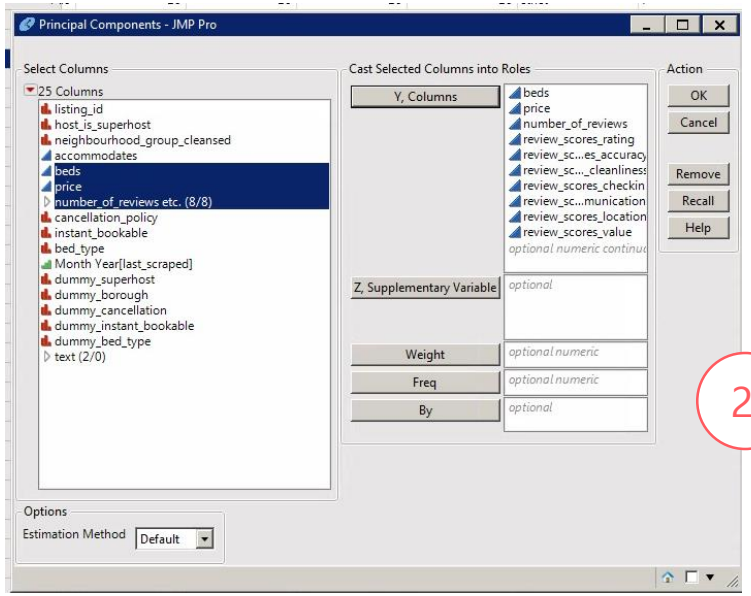
Master1+dummy - Multivariate - JMP Pro

Multivariate

Correlations

	beds	price	number_of_reviews	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value
beds	1.0000	0.2922	0.0896	-0.0366	-0.0508	0.0106	-0.0164	-0.0389	-0.0436	-0.0423
price	0.2922	1.0000	-0.0192	0.0333	0.0075	0.0326	-0.0108	-0.0067	0.1010	-0.0099
number_of_reviews	0.0896	-0.0192	1.0000	-0.0148	0.0005	0.0388	0.0389	0.0267	-0.0469	-0.0231
review_scores_rating	-0.0366	0.0333	-0.0148	1.0000	0.7182	0.7065	0.5613	0.6023	0.4462	0.7498
review_scores_accuracy	-0.0508	0.0075	0.0005	0.7182	1.0000	0.5950	0.5022	0.5320	0.3725	0.6611
review_scores_cleanliness	0.0106	0.0326	0.0388	0.7065	0.5950	1.0000	0.4301	0.4352	0.3173	0.6103
review_scores_checkin	-0.0164	-0.0108	0.0389	0.5613	0.5022	0.4301	1.0000	0.6459	0.3209	0.4929
review_scores_communication	-0.0389	-0.0067	0.0267	0.6023	0.5320	0.4352	0.6459	1.0000	0.3493	0.5299
review_scores_location	-0.0436	0.1010	-0.0469	0.4462	0.3725	0.3173	0.3209	0.3493	1.0000	0.4389
review_scores_value	-0.0423	-0.0099	-0.0231	0.7498	0.6611	0.6103	0.4929	0.5299	0.4389	1.0000

There are 129368 missing values. The correlations are estimated by Pairwise method.



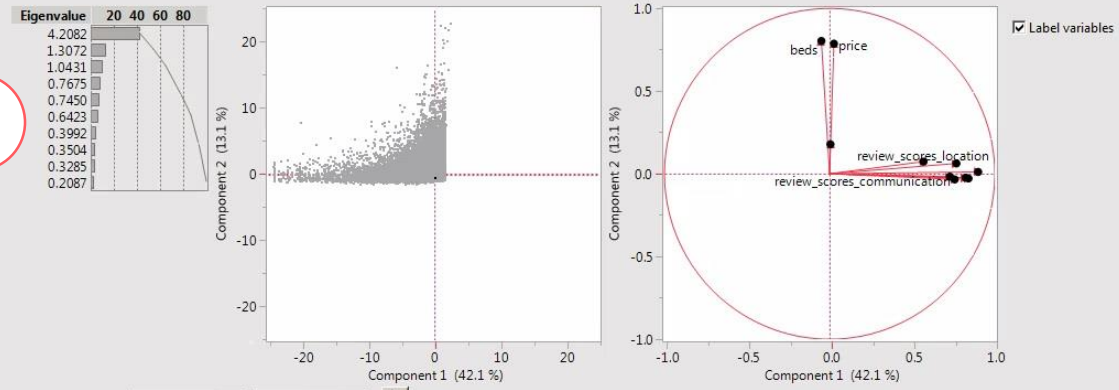
2

3

Master1+dummy - Principal Components - JMP Pro

Principal Components: on Correlations

Summary Plots



Eigenvalues

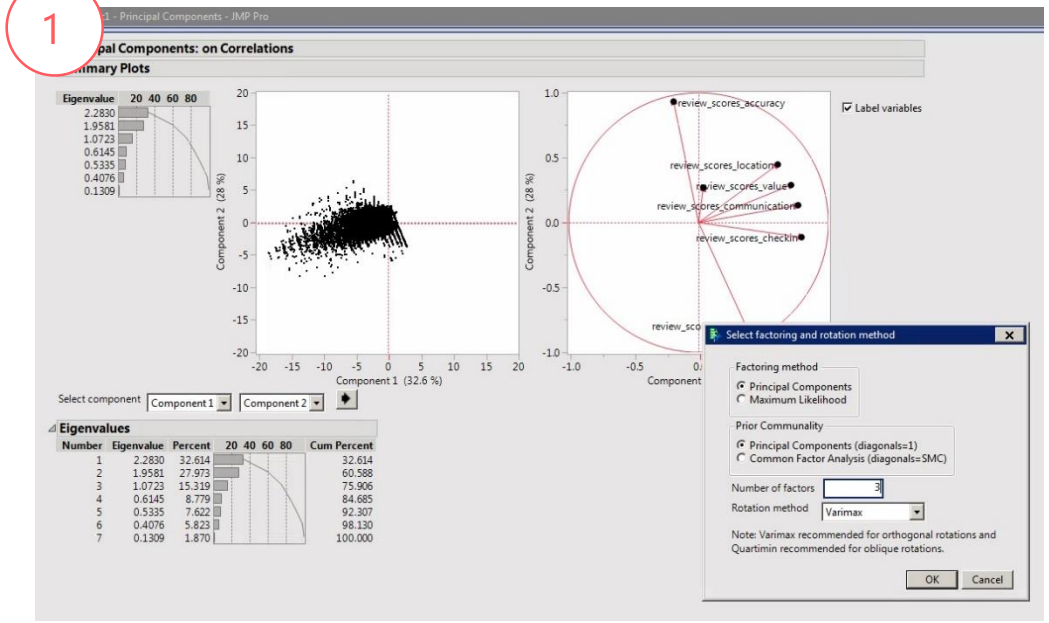
Number	Eigenvalue	Percent	Cum Percent
1	4.2082	42.082	42.082
2	1.3072	13.072	55.154
3	1.0431	10.431	65.584
4	0.7675	7.675	73.259
5	0.7450	7.450	80.709
6	0.6423	6.423	87.131
7	0.3992	3.992	91.123
8	0.3504	3.504	94.628
9	0.3285	3.285	97.913
10	0.2087	2.087	100.000

65.59%

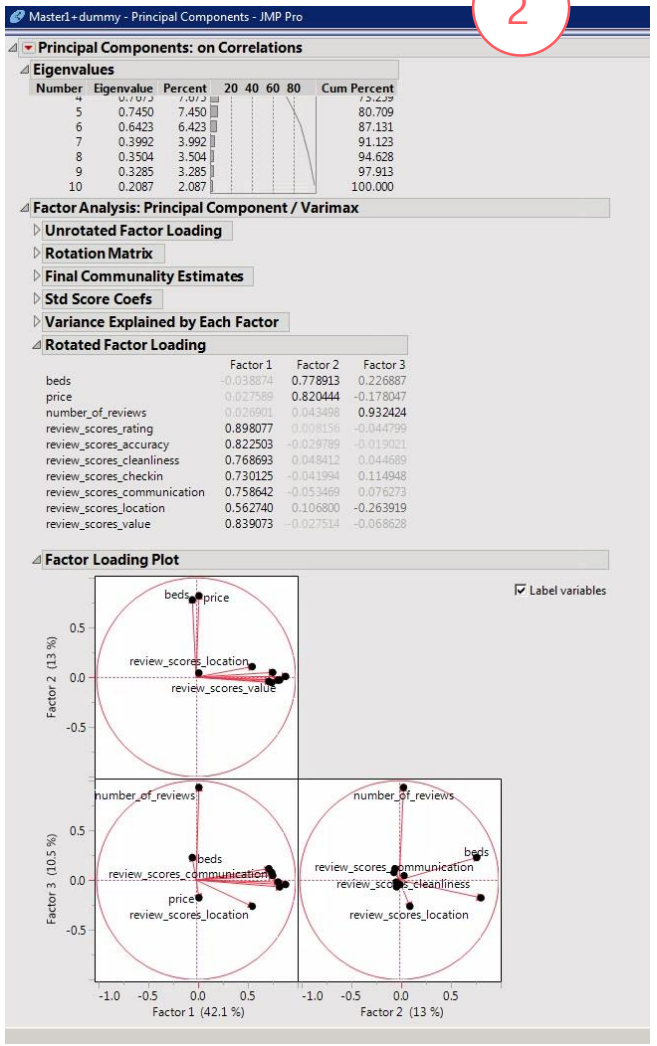
Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Factor Analysis

1



2

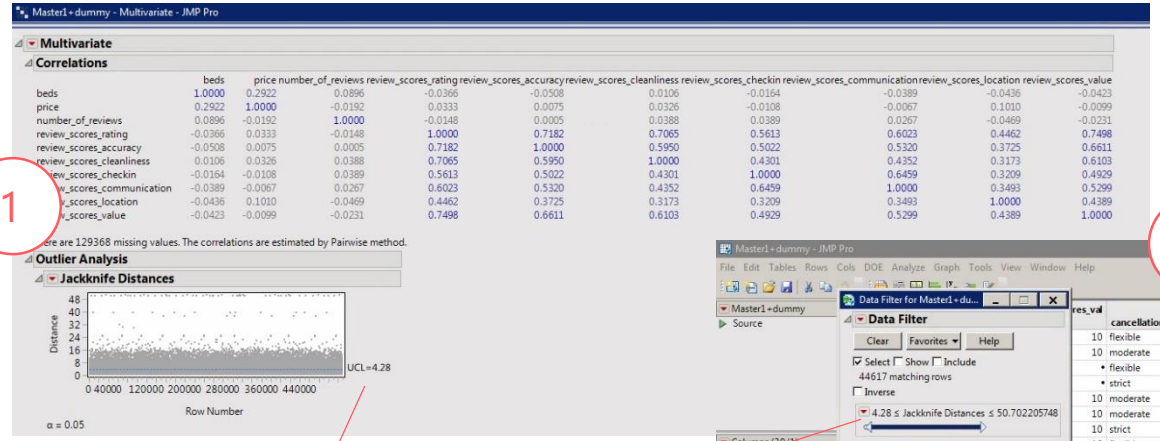


3

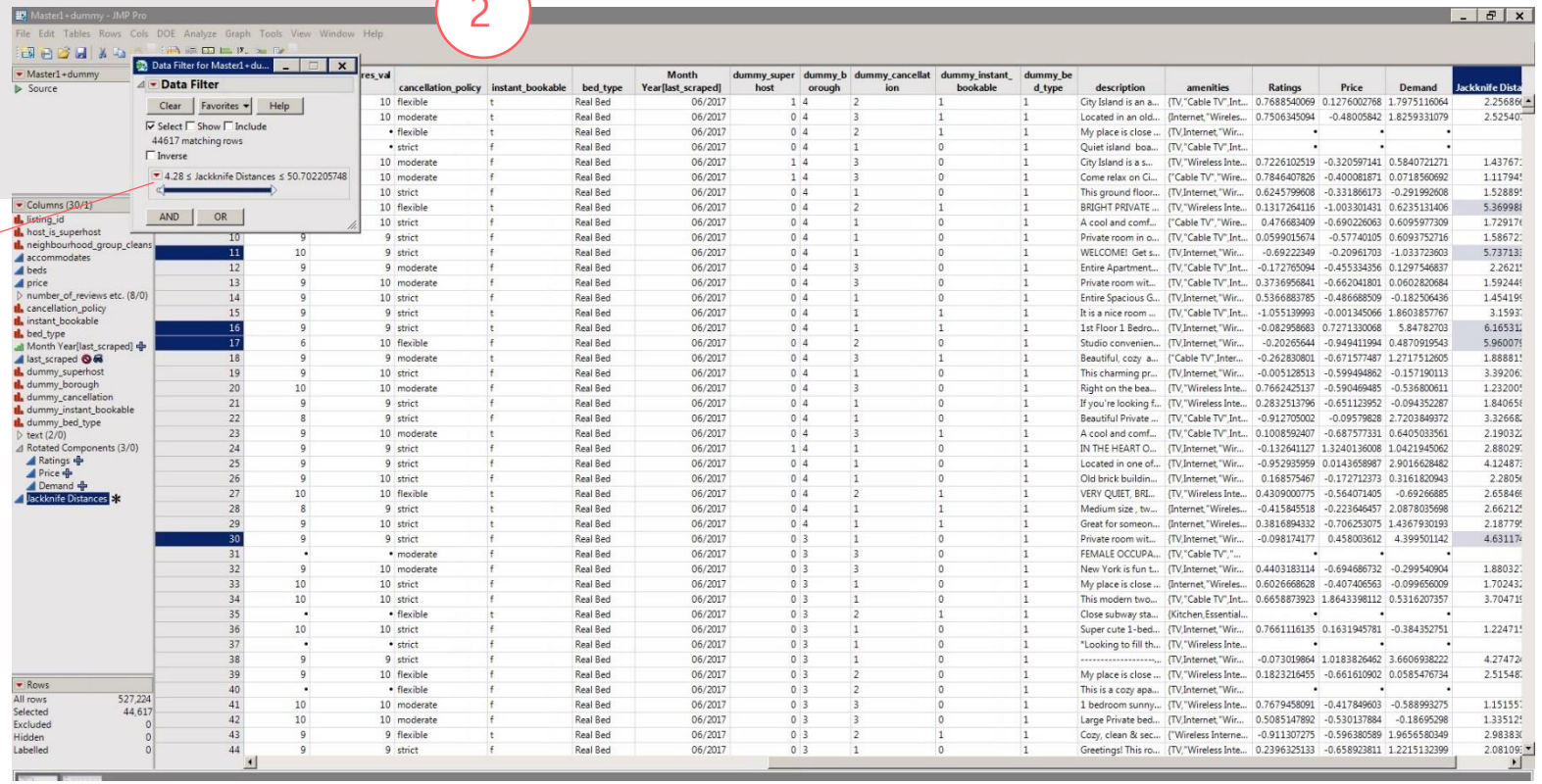
	Ratings	Price	Demand
nt...	0.7688540069	0.1276002768	1.7975116064
s...	0.7506345094	-0.48005842	1.825933107
r...	•	•	
nt...	•	•	
e...	0.7226102519	-0.320597141	0.5840721271
e...	0.7846407826	-0.400081871	0.0718560692
r...	0.6245799608	-0.331866173	-0.291992608
e...	0.1317264116	-1.003301431	0.6235131406
e...	0.476683409	-0.690226063	0.6095977309
nt...	0.0599015674	-0.57740105	0.6093752716
r...	-0.69222349	-0.20961703	-1.033723603
nt...	-0.172765094	-0.455334356	0.1297546837
nt...	0.3736956841	-0.662041801	0.0602820684
r...	0.5366883785	-0.486688509	-0.182506436
nt...	-1.055139993	-0.001345066	1.8603857767
r...	-0.082958683	0.7271330068	5.84782703
r...	-0.20265644	-0.949411994	0.4870919543
...	-0.262830801	-0.671577487	1.2717512605
r...	-0.005128513	-0.599494862	-0.157190113
e...	0.7662425137	-0.590469485	-0.536800611
e...	0.2832513796	-0.651123952	-0.094352287
nt...	-0.912705002	-0.09579828	2.7203849372
nt...	0.1008592407	-0.687577331	0.6405033561
r...	-0.132641127	1.3240136008	1.0421945062
r...	-0.952935959	0.0143658987	2.9016628482
r...	0.168575467	-0.172712373	0.3161820943
e...	0.4309000775	-0.564071405	-0.69266885
s...	-0.415845518	-0.223646457	2.0878035698
s...	0.3816894332	-0.706253075	1.4367930193

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Clustering: Outlier Analysis



1

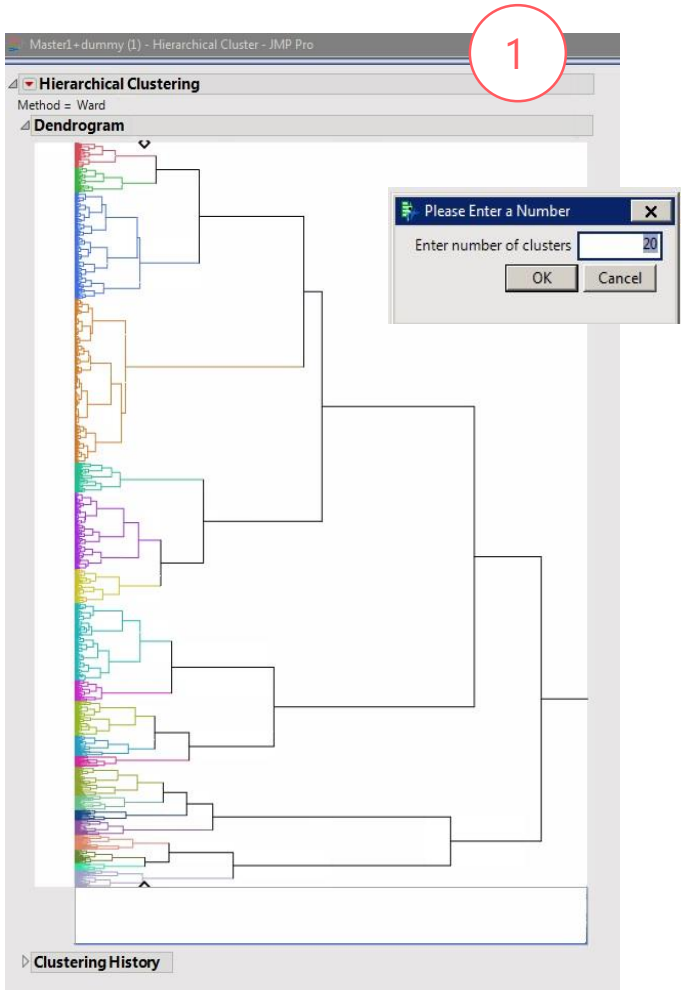


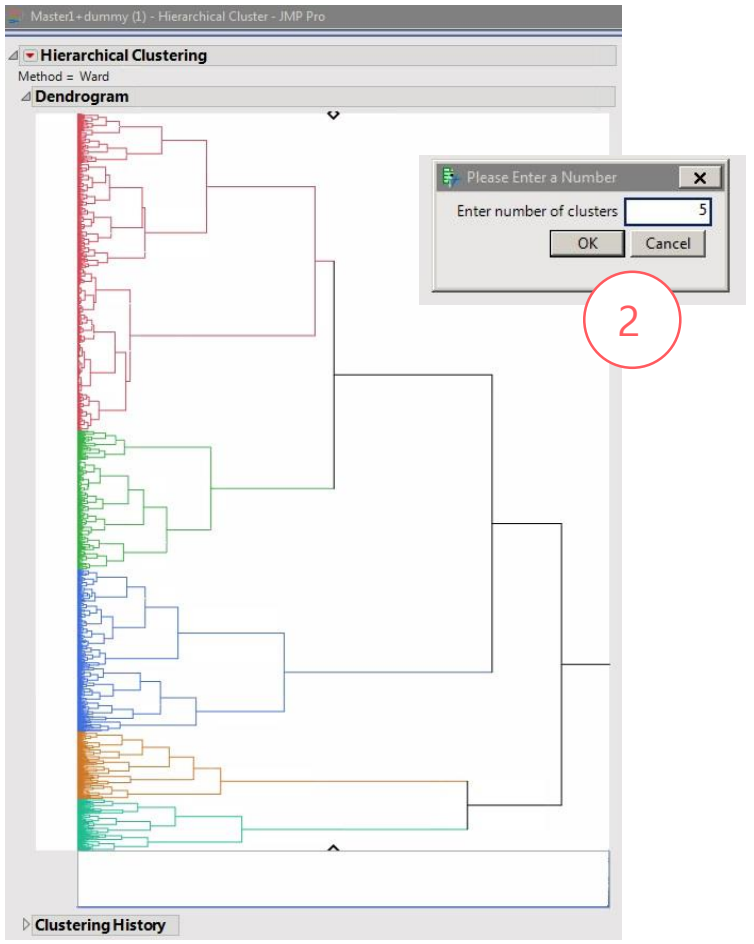
2

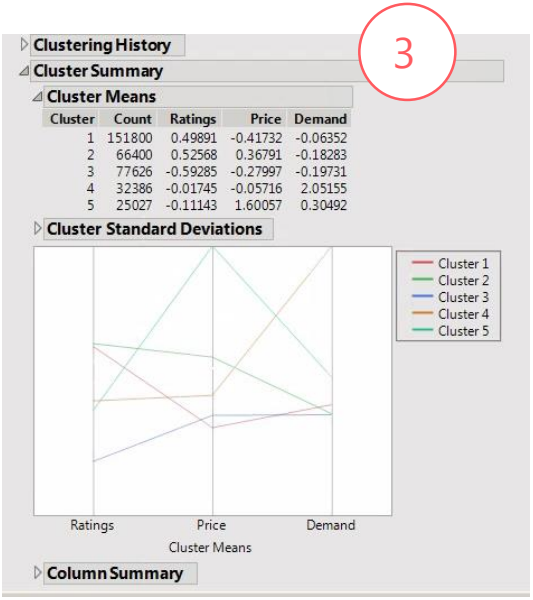
UCL = 4.28

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Hierarchical Clustering by Rotated Components







Problem Overview → **Big Data Issue: Volume** → Analyzing Text Data → Machine Learning → Solution

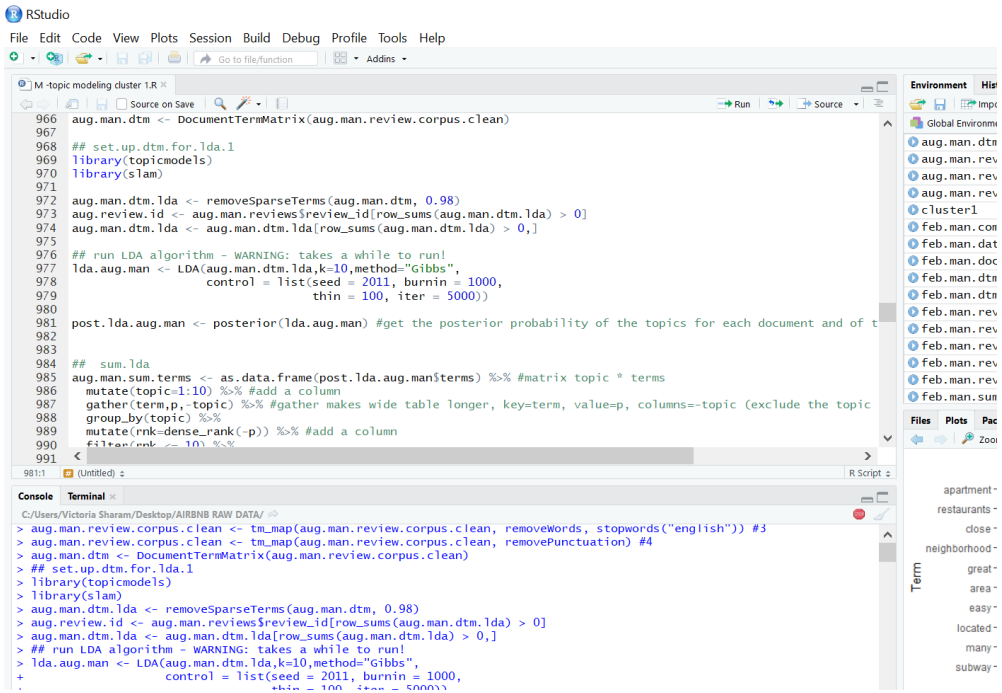
Clusters


1
Mid-Tier
Cost Effective
2
Mid-Tier
3
Low-Tier
4
Conservative
High Demand
5
Luxury

Text Mining: Topic Modeling

1. Joined clusters with review data on JMP

2.



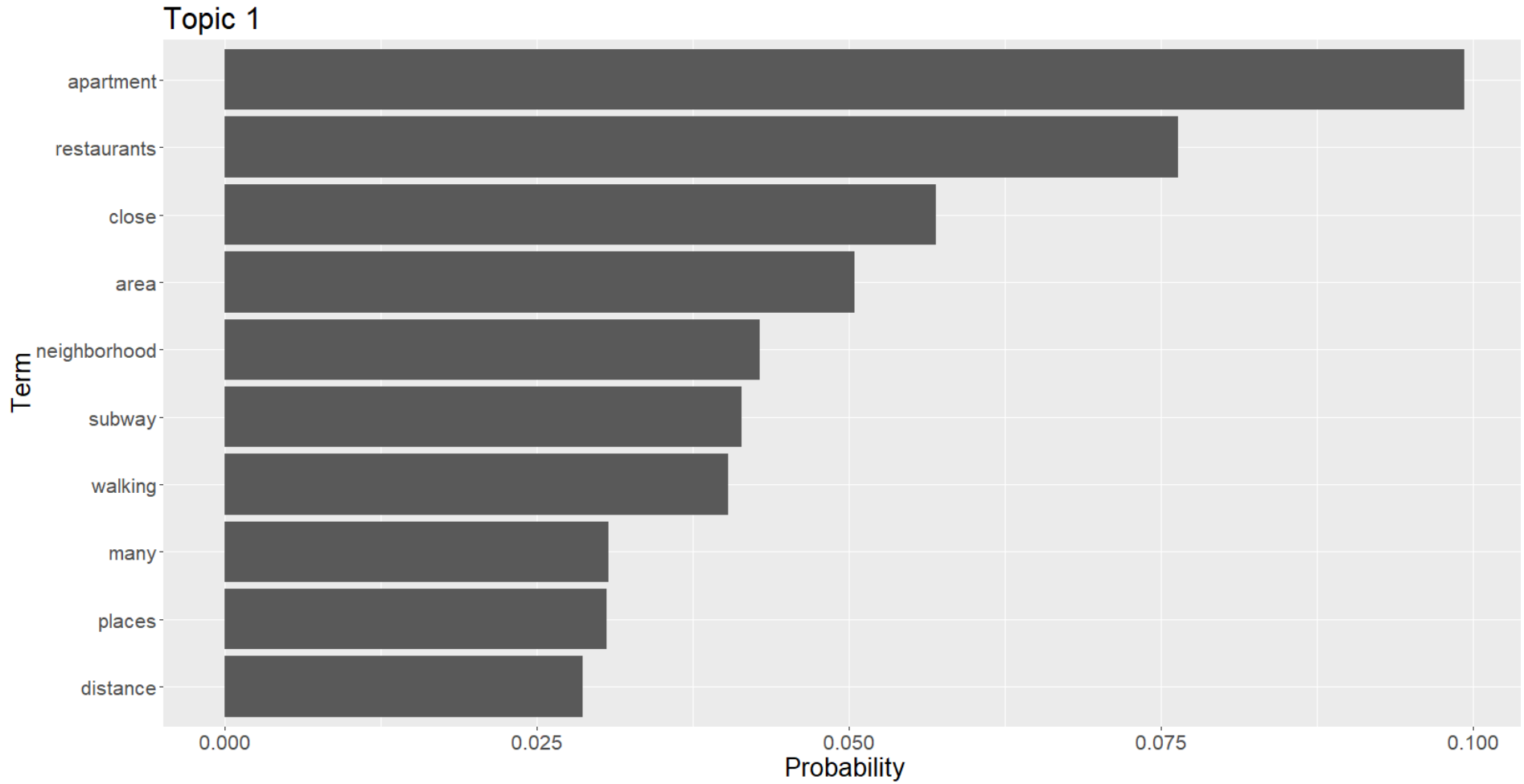
Problem Overview → Big Data Issue: Volume → **Analyzing Text Data** → Machine Learning → Solution

Text Mining: Topic Modeling

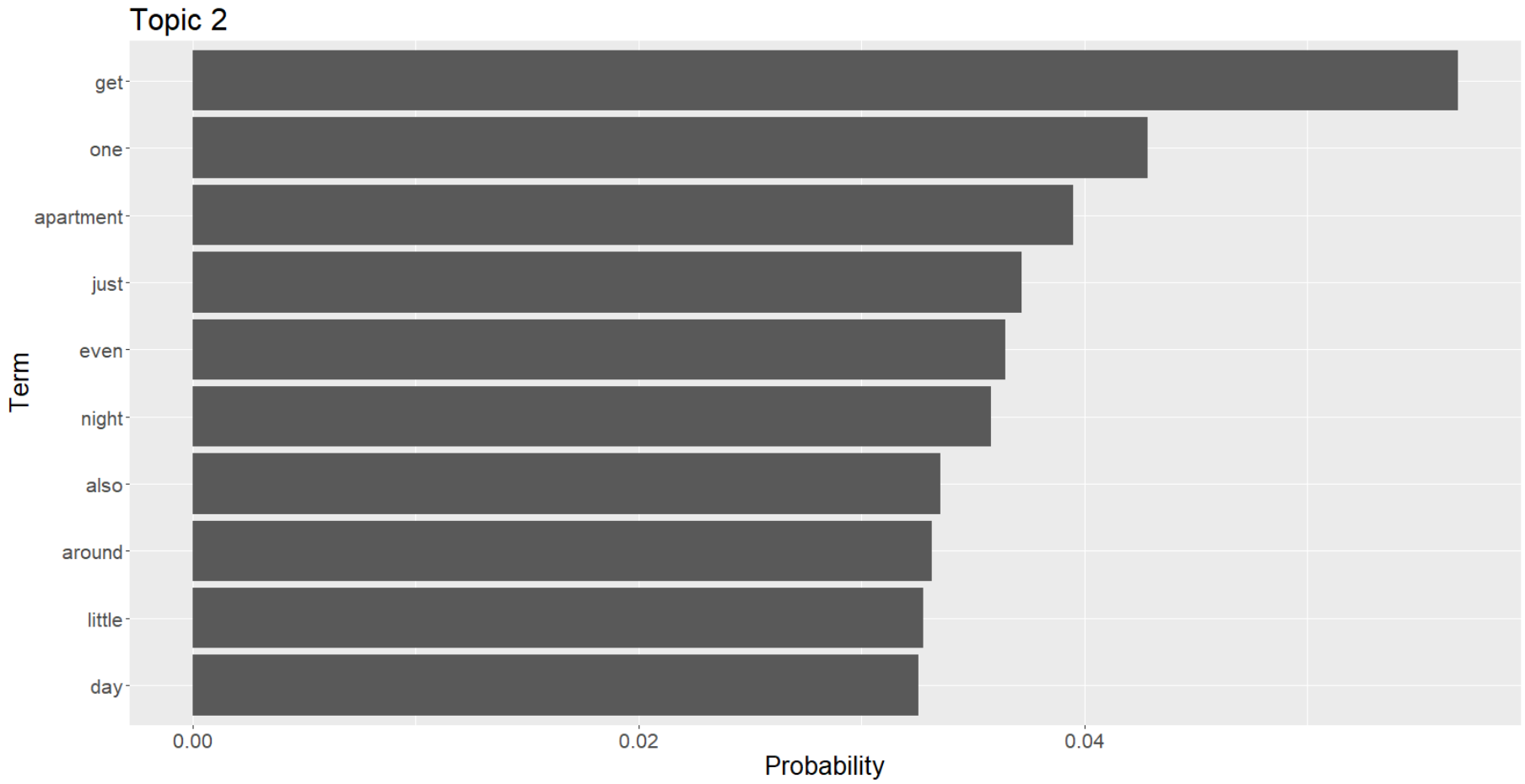
Cluster One, Manhattan, February

Problem Overview → Big Data Issue: Volume → **Analyzing Text Data** → Machine Learning → Solution

Text Mining: Locality



Text Mining: Ephemeral



Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

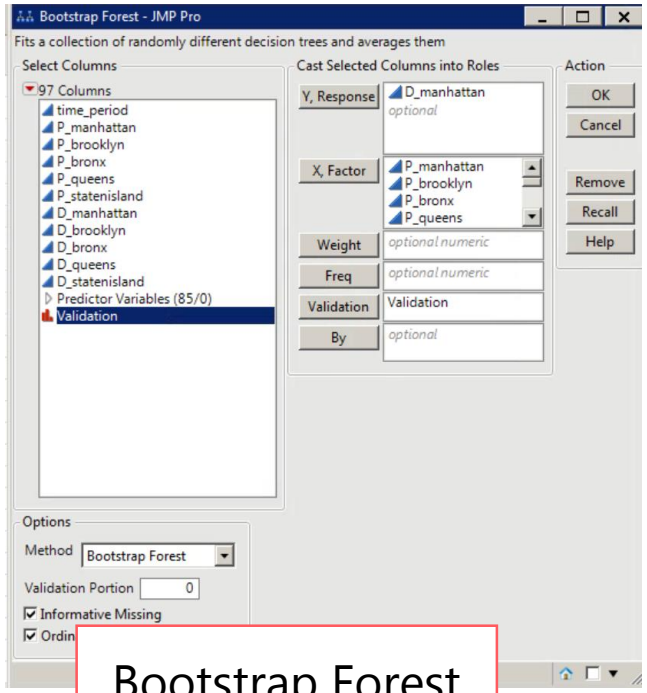
Dataset Structure

	Price	Demand	Superhost	Bed Type	Accommodates	Beds	Cancellation Policy	Review Score Rating	Instant Bookable	Topic Modeling
3/17										
4/17										
5/17										
6/17										
7/17										
8/17										
9/17										
10/17										
11/17										
12/17										
01/18										
02/19										

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

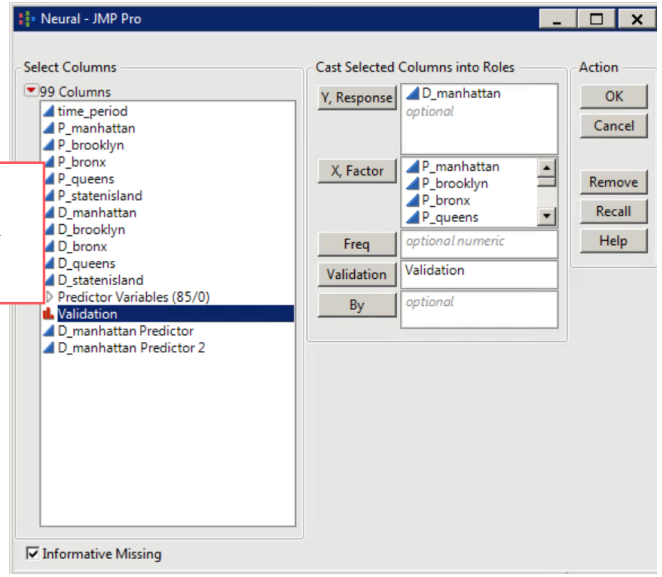
Machine Learning

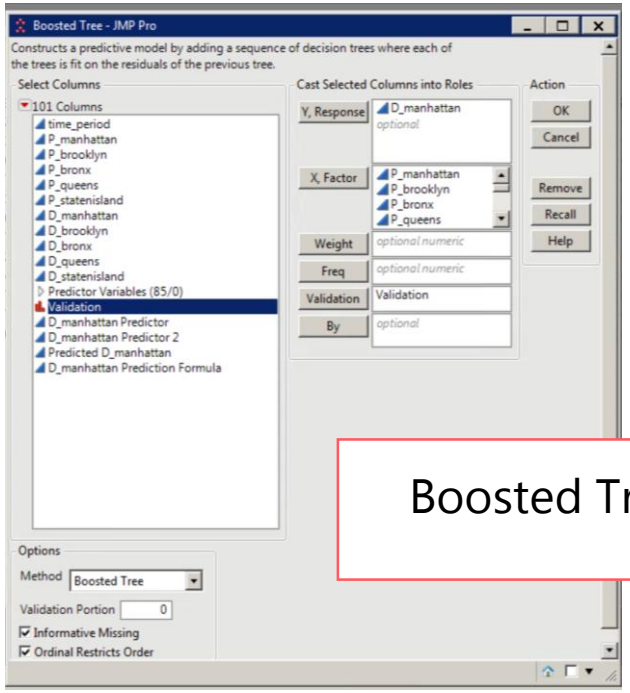
Cluster One, Manhattan



Bootstrap Forest

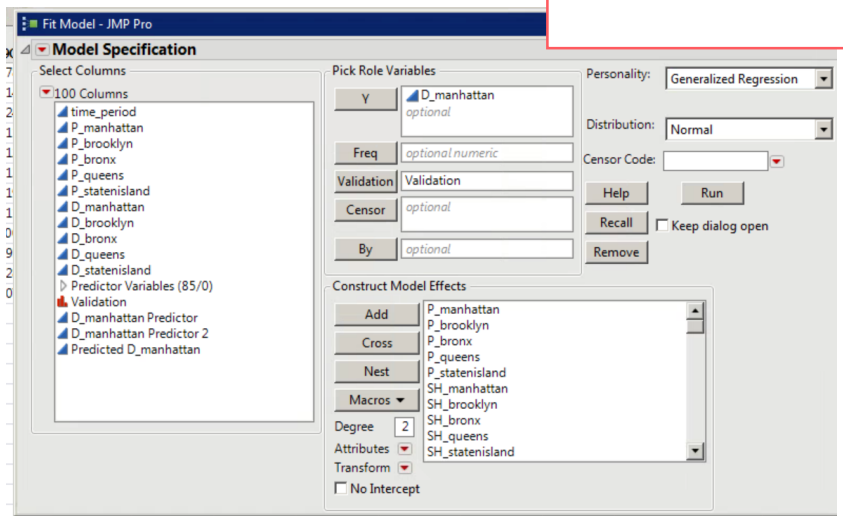
Neural Network





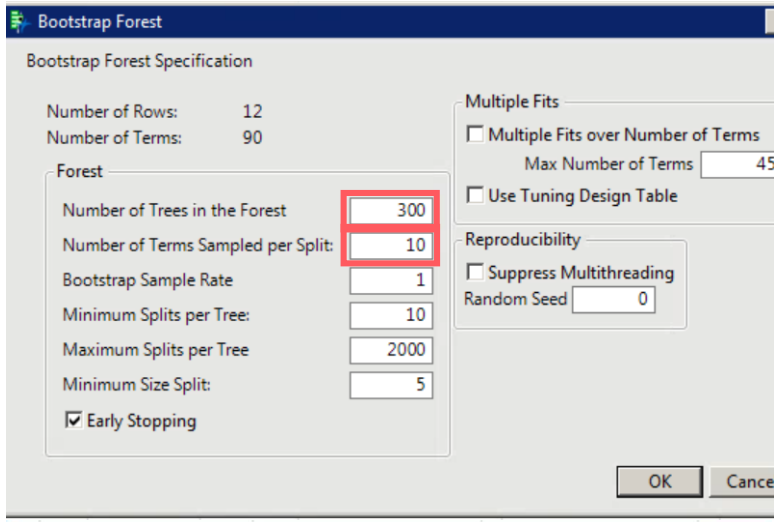
Boosted Tree

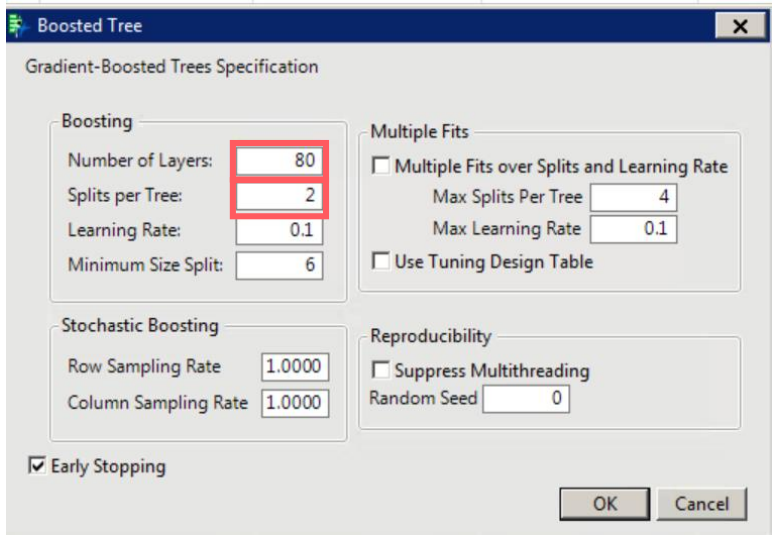
Elastic Net

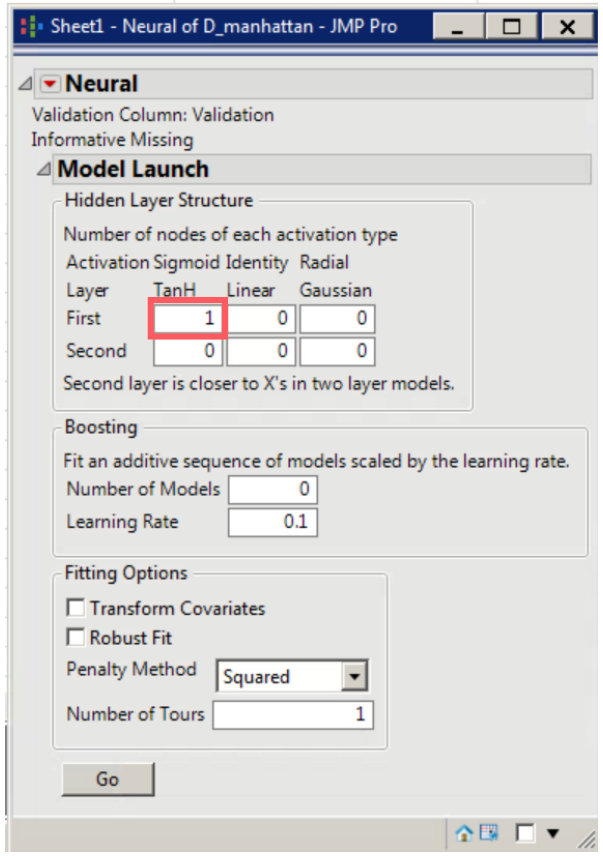


Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

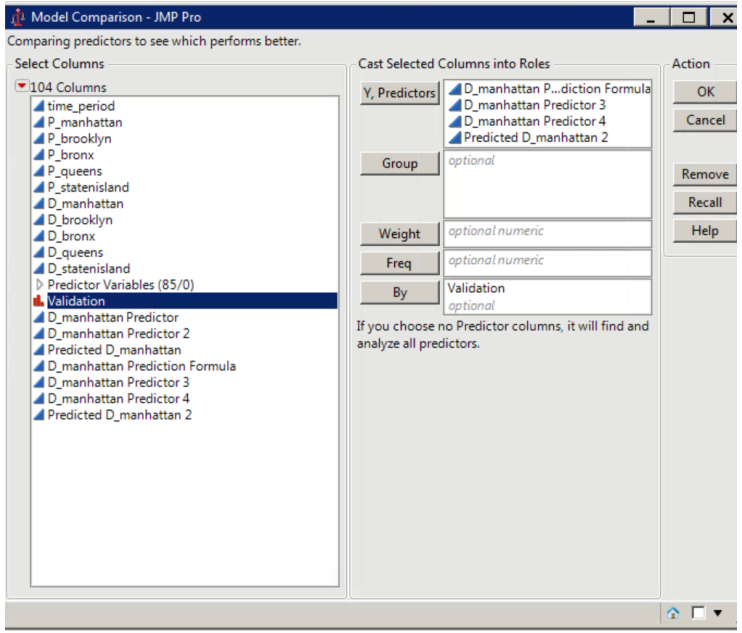
Machine Learning with Tuned Parameters







Model Comparison



Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Machine Learning Model Comparison: RASE

Without Tuning

With Tuning

Sheet1 - Model Comparison - JMP Pro

Model Comparison Validation=Training

Predictors

Measures of Fit for D_manhattan

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_manhattan Predictor	Bootstrap Forest		-0.017	0.4847	0.4079	8
D_manhattan Predictor 2	Boosted Tree		0.0000	0.4806	0.4079	8
Predicted D_manhattan	Neural		0.8277	0.1995	0.1655	8
D_manhattan Prediction Formula	Fit Generalized Adaptive Elastic Net		0.9237	0.1327	0.1078	8

Model Comparison Validation=Validation

Predictors

Measures of Fit for D_manhattan

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_manhattan Predictor	Bootstrap Forest		-0.023	0.4833	0.4253	4
D_manhattan Predictor 2	Boosted Tree		-0.080	0.4965	0.4566	4
Predicted D_manhattan	Neural		0.9575	0.0985	0.0951	4
D_manhattan Prediction Formula	Fit Generalized Adaptive Elastic Net		0.8082	0.2092	0.1740	4

Sheet1 - Model Comparison - JMP Pro

Model Comparison Validation=Training

Predictors

Measures of Fit for D_manhattan

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_manhattan Prediction Formula	Fit Generalized Adaptive Elastic Net		0.9237	0.1327	0.1078	8
D_manhattan Predictor 3	Bootstrap Forest		-0.091	0.5020	0.4079	8
D_manhattan Predictor 4	Boosted Tree		0.0000	0.4806	0.4079	8
Predicted D_manhattan 2	Neural		0.9424	0.1153	0.0897	8

Model Comparison Validation=Validation

Predictors

Measures of Fit for D_manhattan

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_manhattan Prediction Formula	Fit Generalized Adaptive Elastic Net		0.8082	0.2092	0.1740	4
D_manhattan Predictor 3	Bootstrap Forest		-0.000	0.4779	0.3841	4
D_manhattan Predictor 4	Boosted Tree		-0.080	0.4965	0.4566	4
Predicted D_manhattan 2	Neural		0.9436	0.1135	0.0845	4

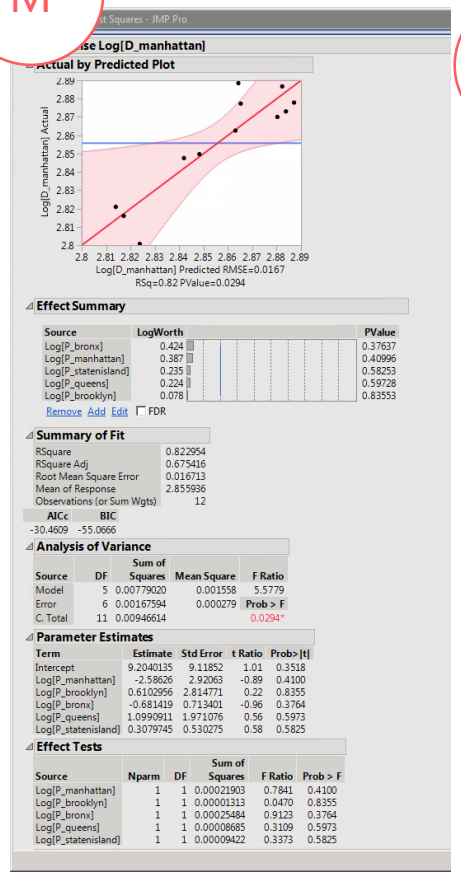
Neural Network

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Optimal Pricing for Airbnb Hosts

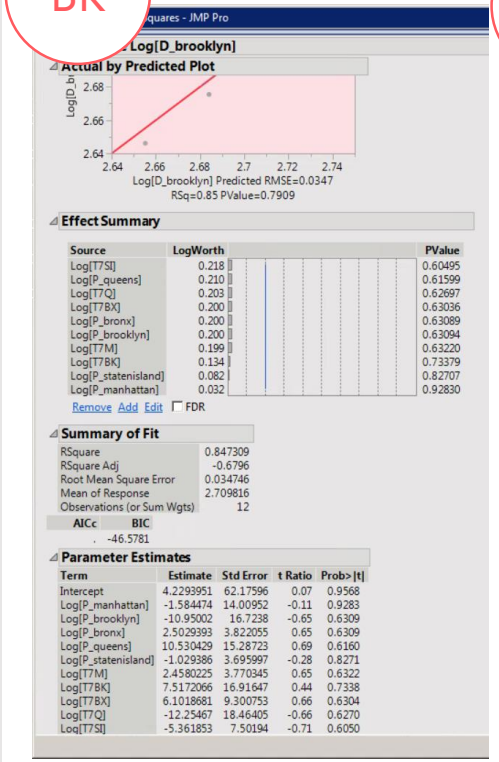
M

Log-Log



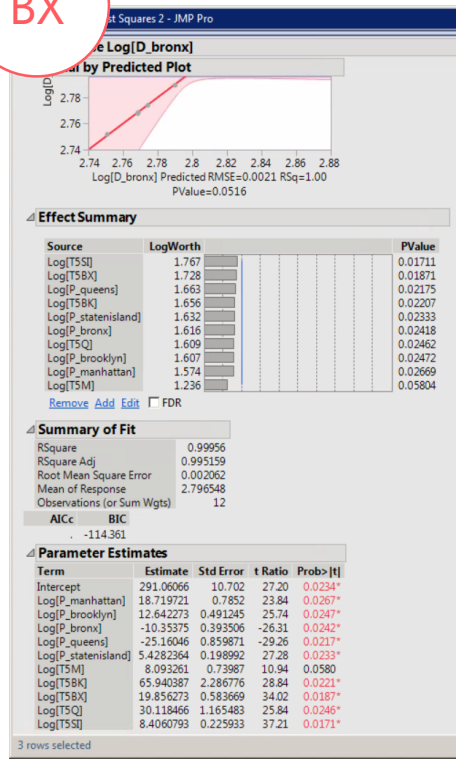
BK

Log-Log



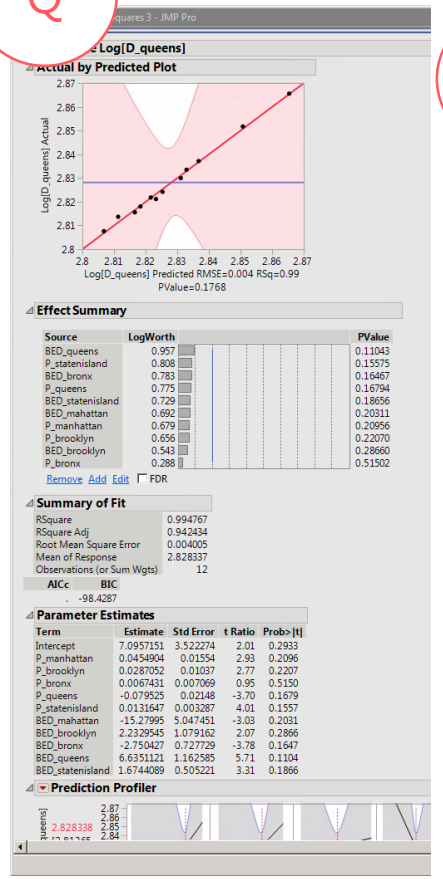
BX

Log-Log



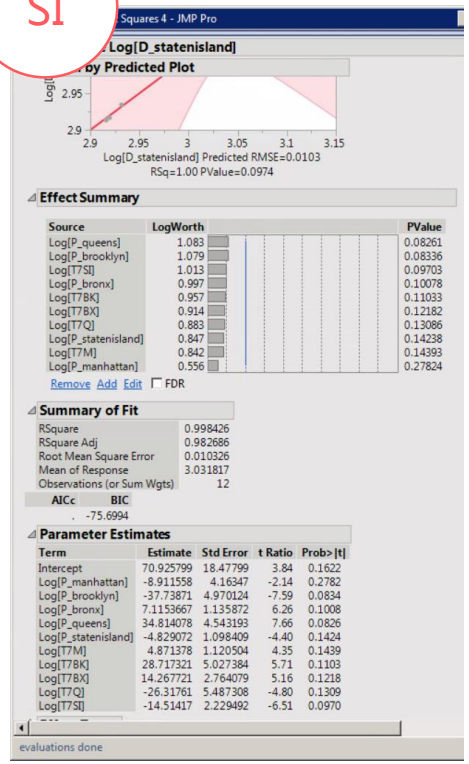
Q

Semi-Log



SI

Log-Log



Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Manhattan Optimal Daily Price

$220 / day

Based on this price, a Manhattan host would have to host a guest for **19 days** to make their monthly rent.

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Brooklyn Optimal Daily Price

airbnb

$112 / day

Based on this price, a Brooklyn host would have to host a guest for **28 days** to make their monthly rent.

Problem Overview

Big Data Issue: Volume

Analyzing Text Data

Machine Learning

Solution

Bronx Optimal Daily Price

$63 / day

Based on this price, a Bronx host would have to host a guest for **27 days** to make their monthly rent.

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Queens Optimal Daily Price

$120 / day

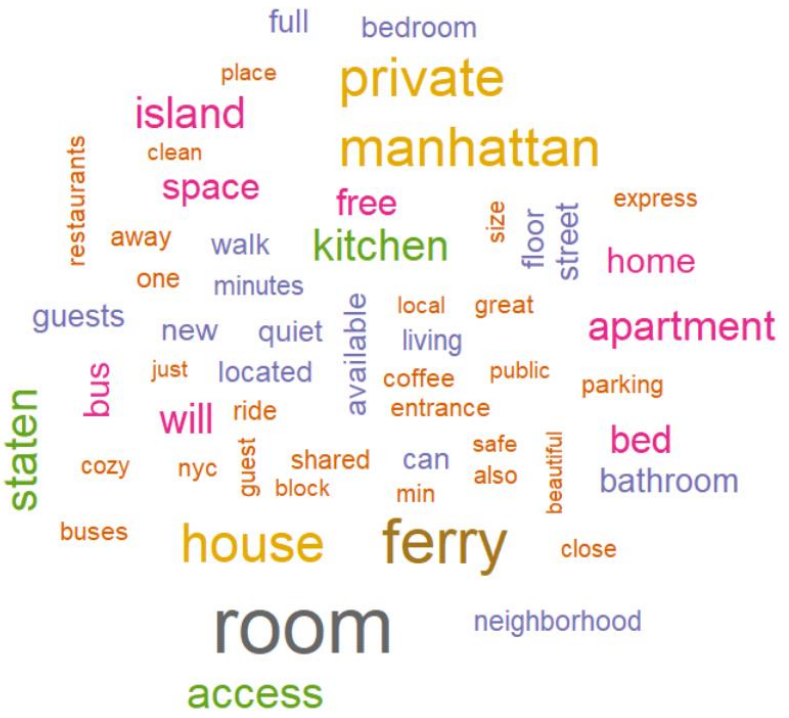
Based on this price, a Queens host would have to host a guest for **25 days** to make their monthly rent.

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Staten Island Optimal Daily Price

$63 / day

Based on this price, a Staten Island host would have to host a guest for **24 days** to make their monthly rent.



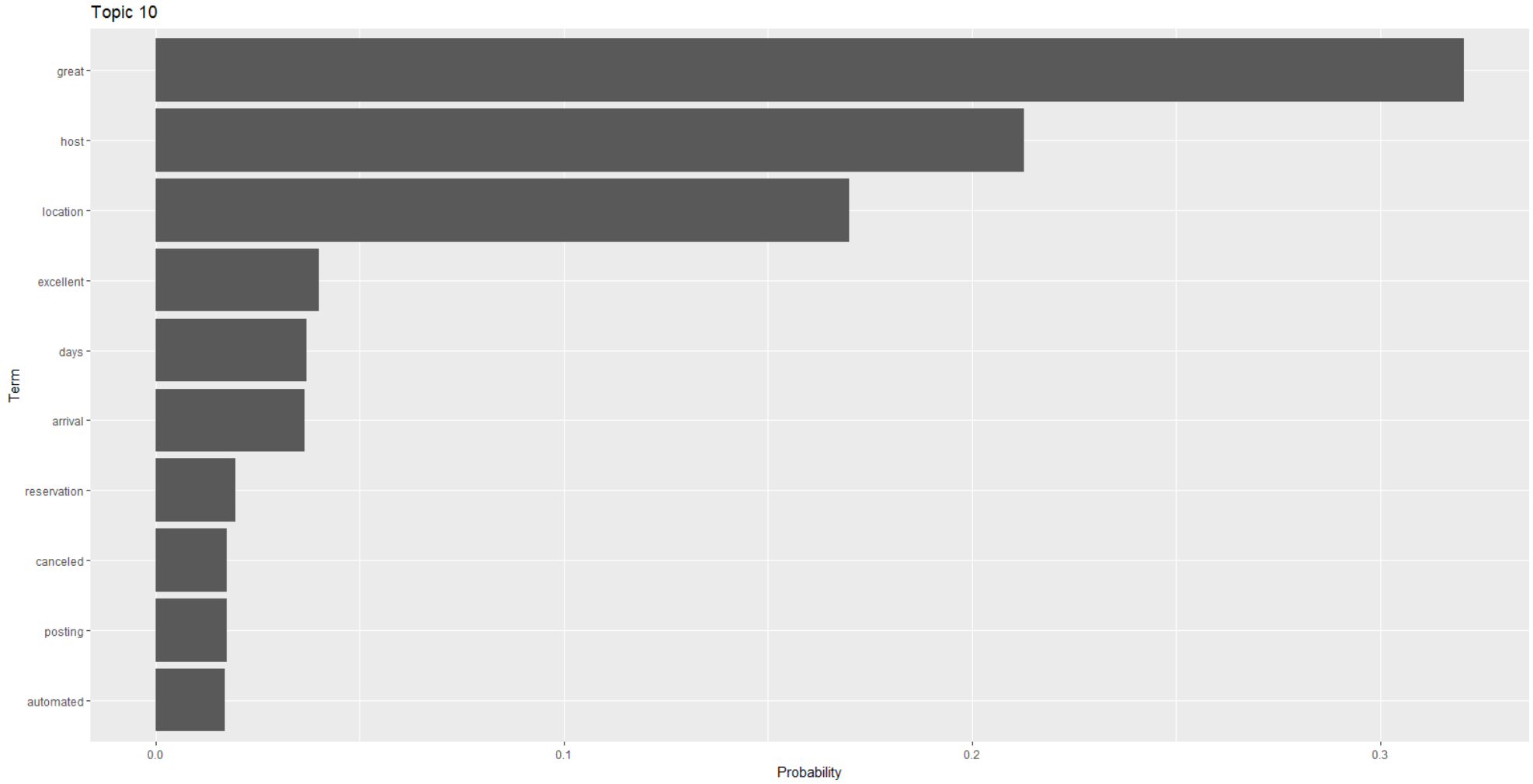
Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Thank you!

Any questions?

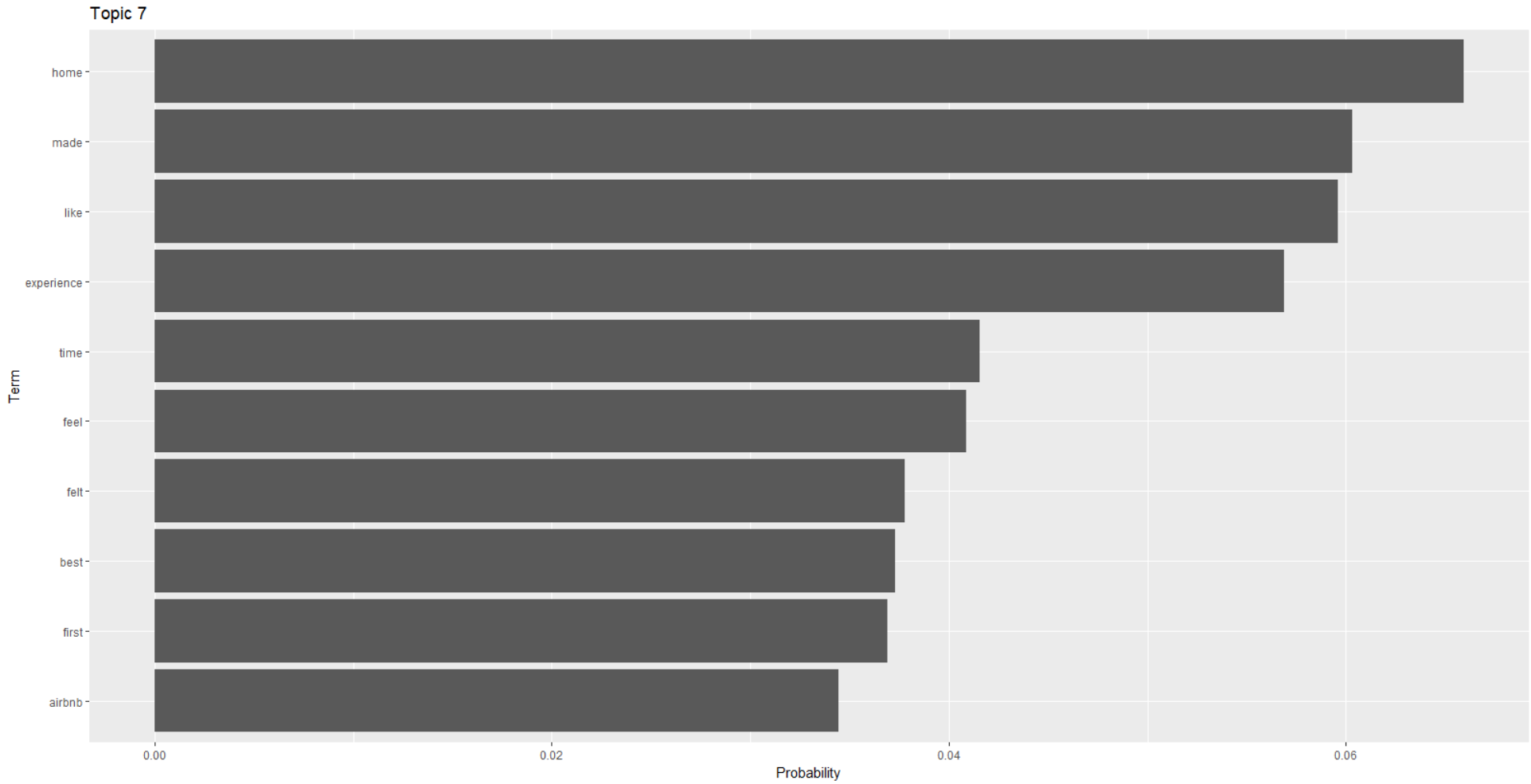
Appendix

Text Mining: Ease of Booking

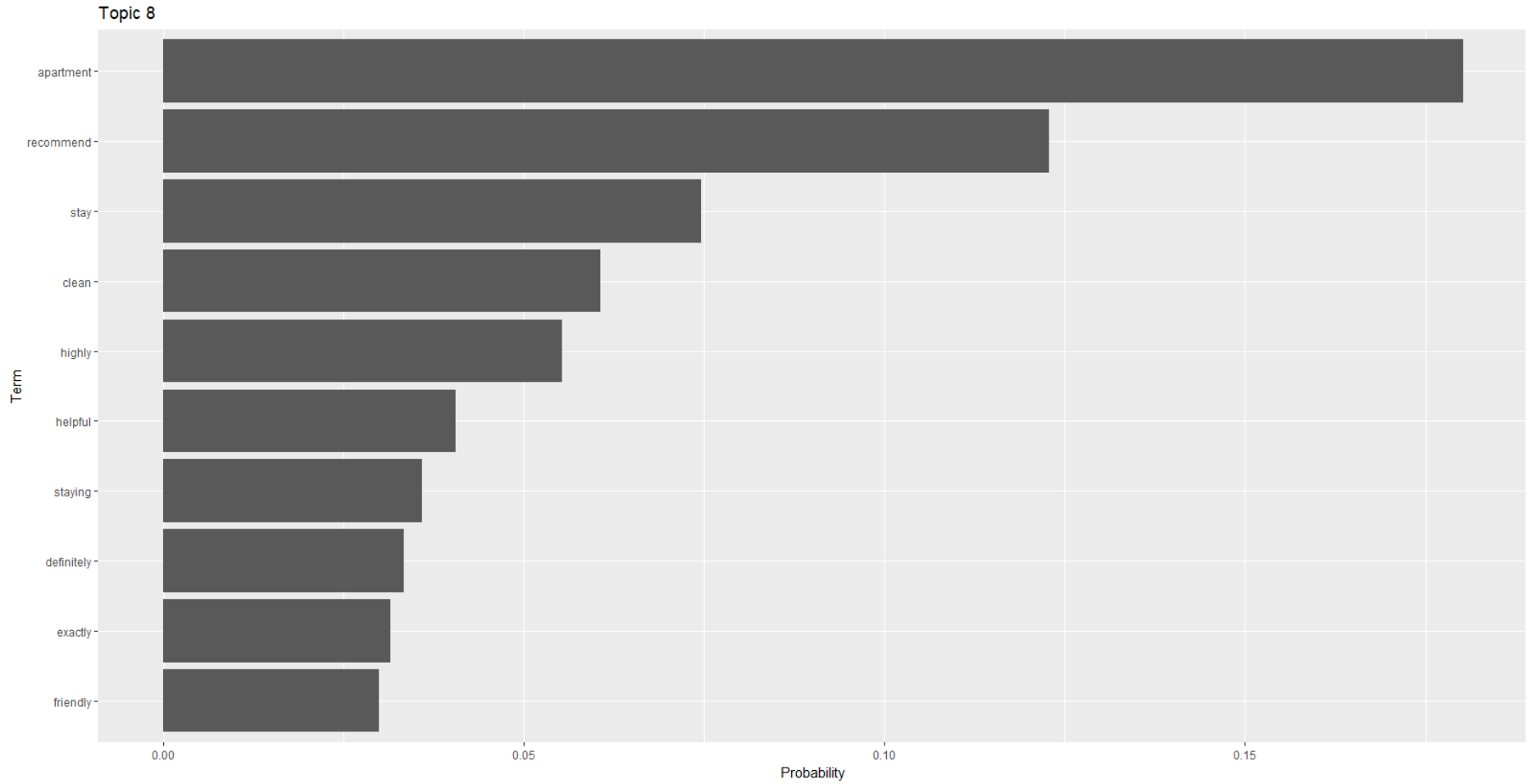


Problem Overview → Big Data Issue: Volume → **Analyzing Text Data** → Machine Learning → Solution

Text Mining: Satisfaction

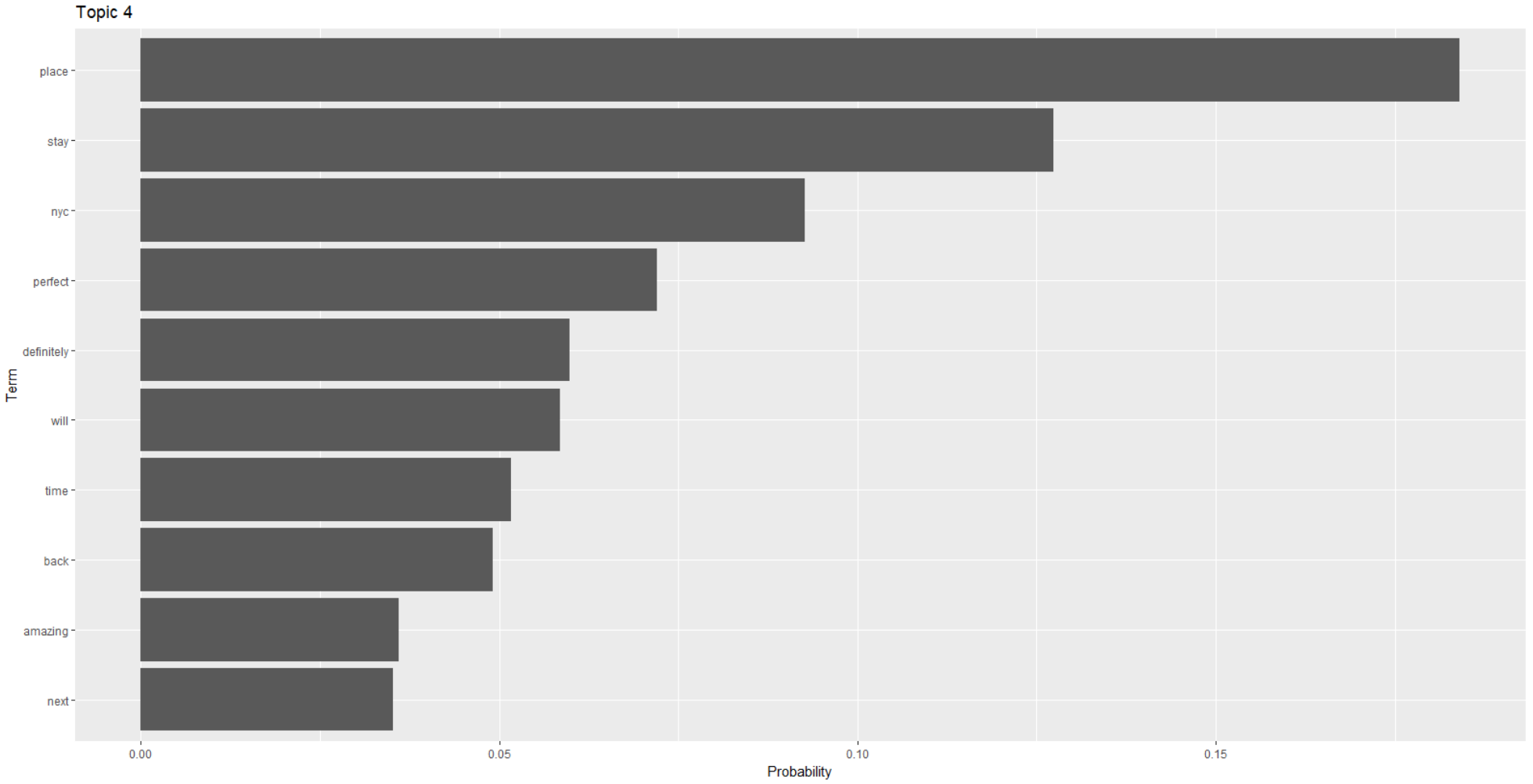


Text Mining: Would Recommend



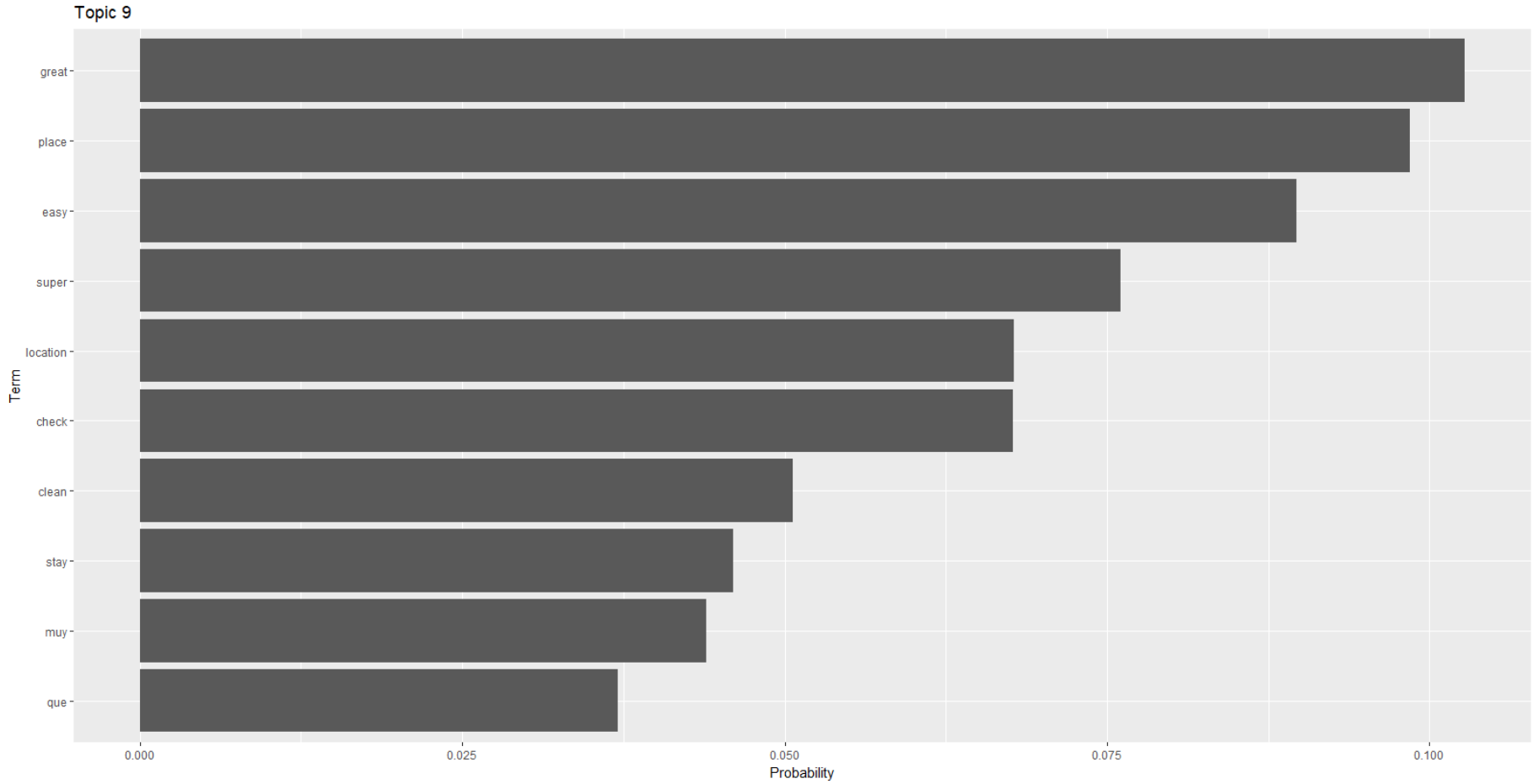
Problem Overview → Big Data Issue: Volume → **Analyzing Text Data** → Machine Learning → Solution

Text Mining: Satisfaction II



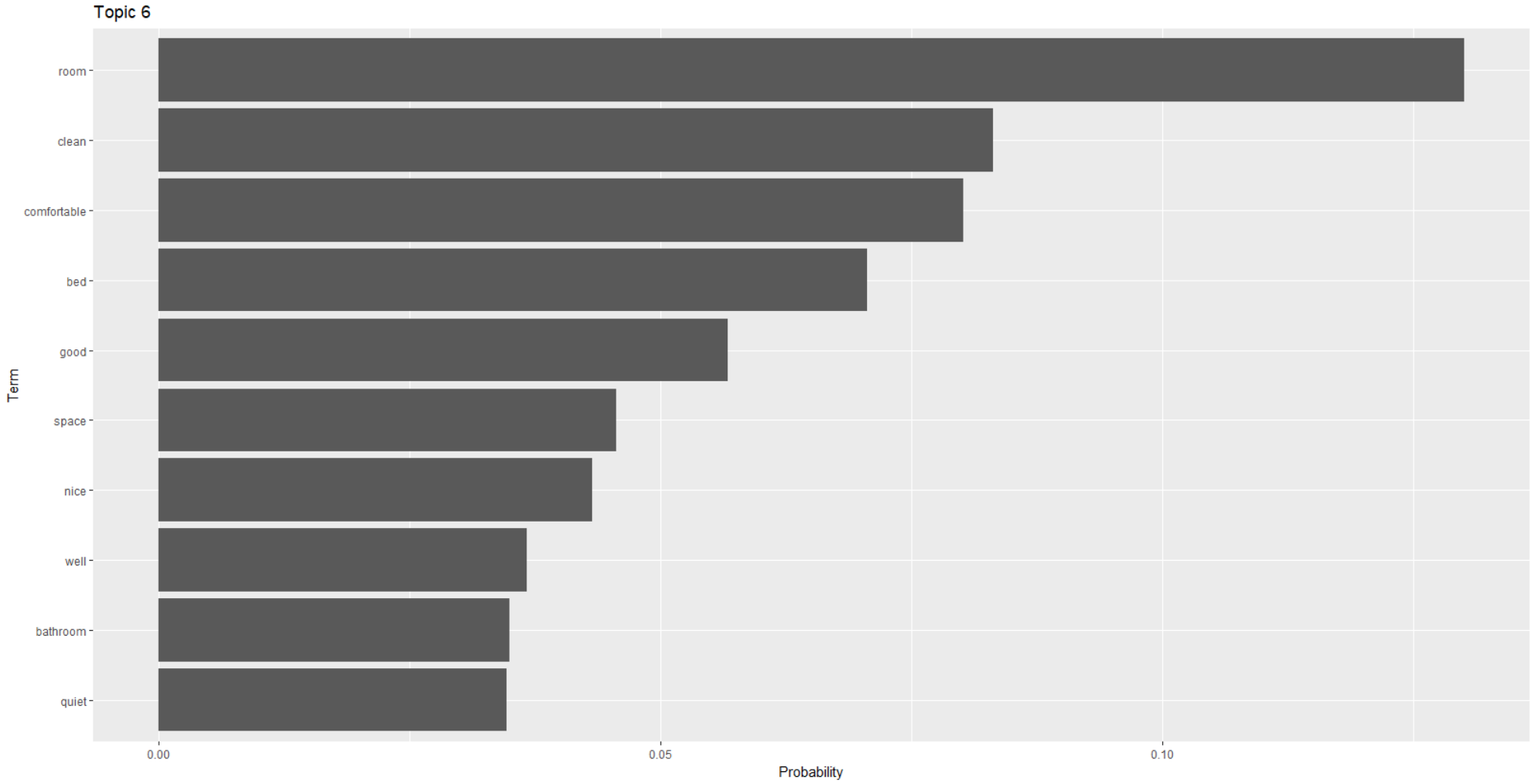
Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Text Mining: Overall Experience

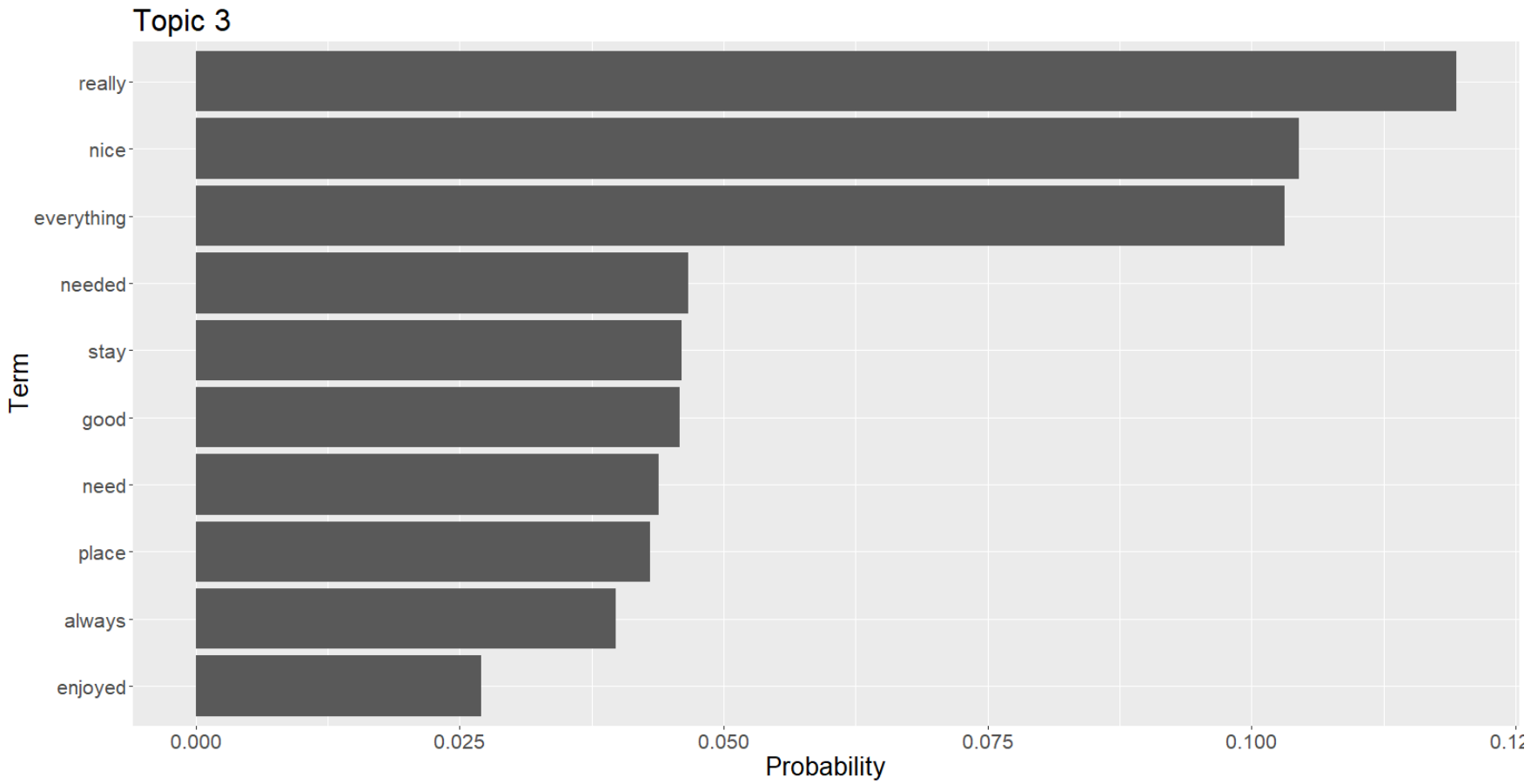


Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

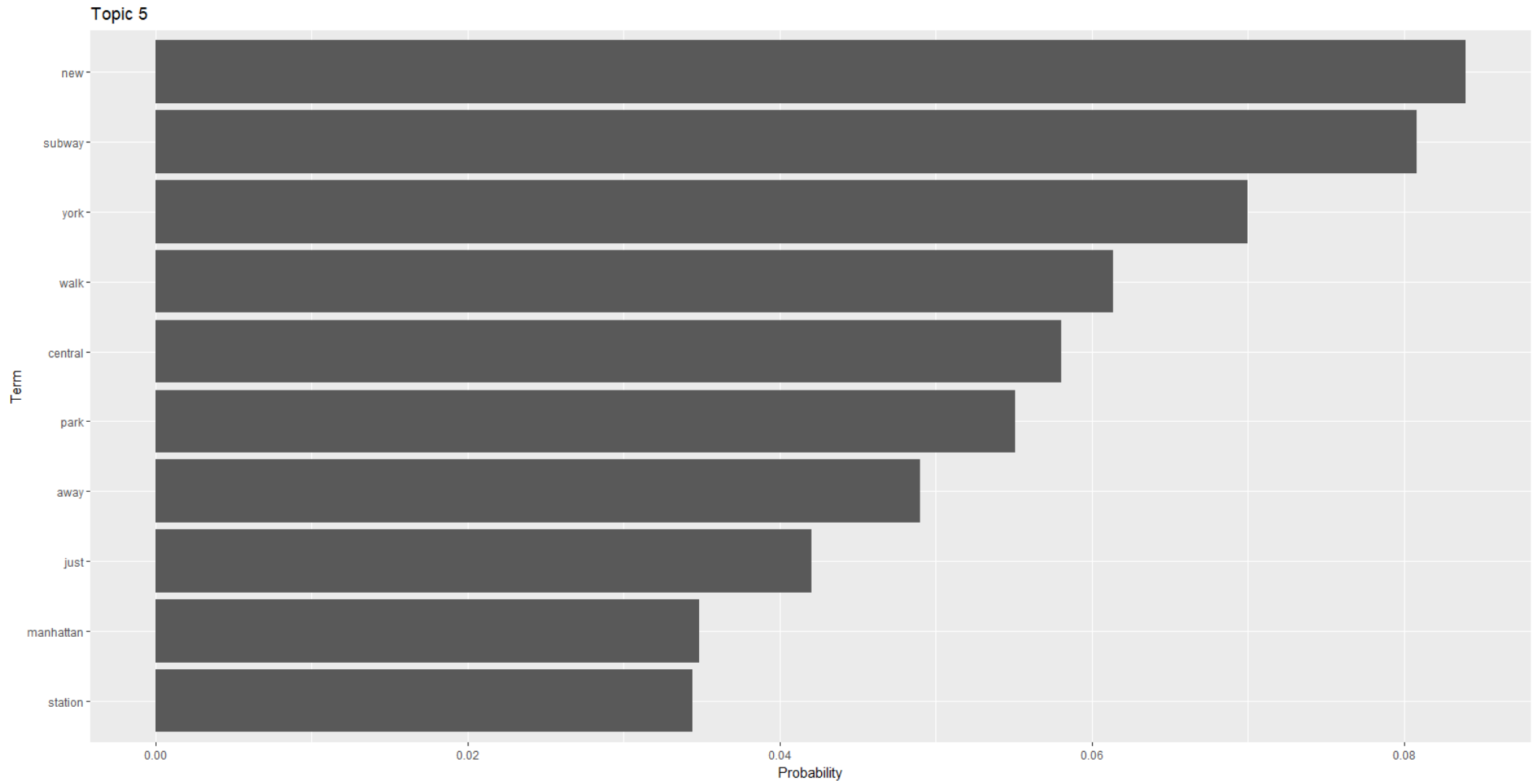
Text Mining: Quality



Text Mining: Enjoyment



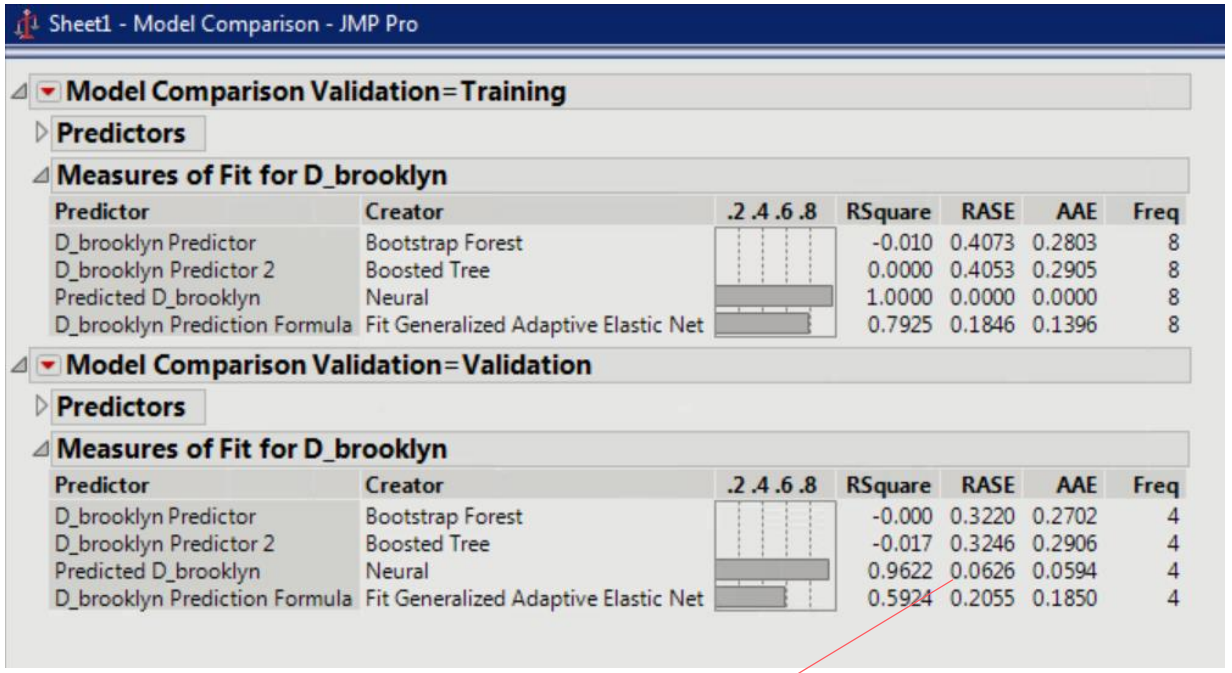
Text Mining: Location Perks



Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Machine Learning

Cluster One, Brooklyn



Neural Network

Machine Learning

Cluster One, Bronx

Sheet1 - Model Comparison - JMP Pro

Model Comparison Validation=Training

Predictors

Measures of Fit for D_bronx

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_bronx Predictor	Bootstrap Forest		-0.064	0.5390	0.4019	8
D_bronx Predictor 2	Boosted Tree		0.0000	0.5226	0.3981	8
Predicted D_bronx	Neural		1.0000	0.0018	0.0015	8
D_bronx Prediction Formula	Fit Generalized Adaptive Elastic Net		0.7718	0.2496	0.1890	8

Model Comparison Validation=Validation

Predictors

Measures of Fit for D_bronx

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_bronx Predictor	Bootstrap Forest		-0.001	0.3176	0.2386	4
D_bronx Predictor 2	Boosted Tree		-0.200	0.3478	0.3047	4
Predicted D_bronx	Neural		0.1382	0.2947	0.2529	4
D_bronx Prediction Formula	Fit Generalized Adaptive Elastic Net		0.7188	0.1683	0.1490	4

Elastic Net

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Machine Learning

Cluster One, Queens

Sheet1 - Model Comparison - JMP Pro

Model Comparison Validation=Training

Predictors

Measures of Fit for D_queens

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_queens Predictor	Bootstrap Forest		-0.014	0.2855	0.2140	8
D_queens Predictor 2	Boosted Tree		0.0000	0.2836	0.2140	8
Predicted D_queens	Neural		0.2751	0.2414	0.2159	8
D_queens Prediction Formula	Fit Generalized Adaptive Elastic Net		0.7565	0.1399	0.1039	8

Model Comparison Validation=Validation

Predictors

Measures of Fit for D_queens

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_queens Predictor	Bootstrap Forest		-0.000	0.2479	0.2091	4
D_queens Predictor 2	Boosted Tree		-0.017	0.2500	0.2257	4
Predicted D_queens	Neural		0.8544	0.0946	0.0730	4
D_queens Prediction Formula	Fit Generalized Adaptive Elastic Net		0.6258	0.1517	0.1173	4

Neural Network

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution

Machine Learning

Cluster One, Staten Island

Sheet1 - Model Comparison - JMP Pro

Model Comparison Validation=Training

Predictors

Measures of Fit for D_statenisland

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_statenisland Predictor	Bootstrap Forest		-0.180	1.5567	1.2299	8
D_statenisland Predictor 2	Boosted Tree		0.0000	1.4328	1.1686	8
Predicted D_statenisland	Neural		0.6677	0.8259	0.6441	8
D_statenisland Prediction Formula	Fit Generalized Adaptive Elastic Net		0.7760	0.6782	0.5218	8

Model Comparison Validation=Validation

Predictors

Measures of Fit for D_statenisland

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
D_statenisland Predictor	Bootstrap Forest		-0.000	1.6603	1.4152	4
D_statenisland Predictor 2	Boosted Tree		-0.148	1.7789	1.7195	4
Predicted D_statenisland	Neural		0.9322	0.4321	0.3823	4
D_statenisland Prediction Formula	Fit Generalized Adaptive Elastic Net		0.5948	1.0567	1.0341	4

Neural Network

Problem Overview → Big Data Issue: Volume → Analyzing Text Data → Machine Learning → Solution