Hotel Reviews

Text Mining & Sentiment Analysis

Weijun Zhu **Chuan Wu** **Dinglun Tan**

Contents

01 Data Description

02 Pre-processing

03 Visulization&Analytics

04 Future work

PART 01

Data Description

Where

This dataset comes from Kaggle and provided by Datafiniti's Business Database

What

This is list of 1000 hotels with over 30,000 data, and their reviews. The dataset includes the location, name, rating, review data and more

Why

We want to help companies, like Expedia, Airbnb, etc to imporve their recommendation system, and bring good customer experience.

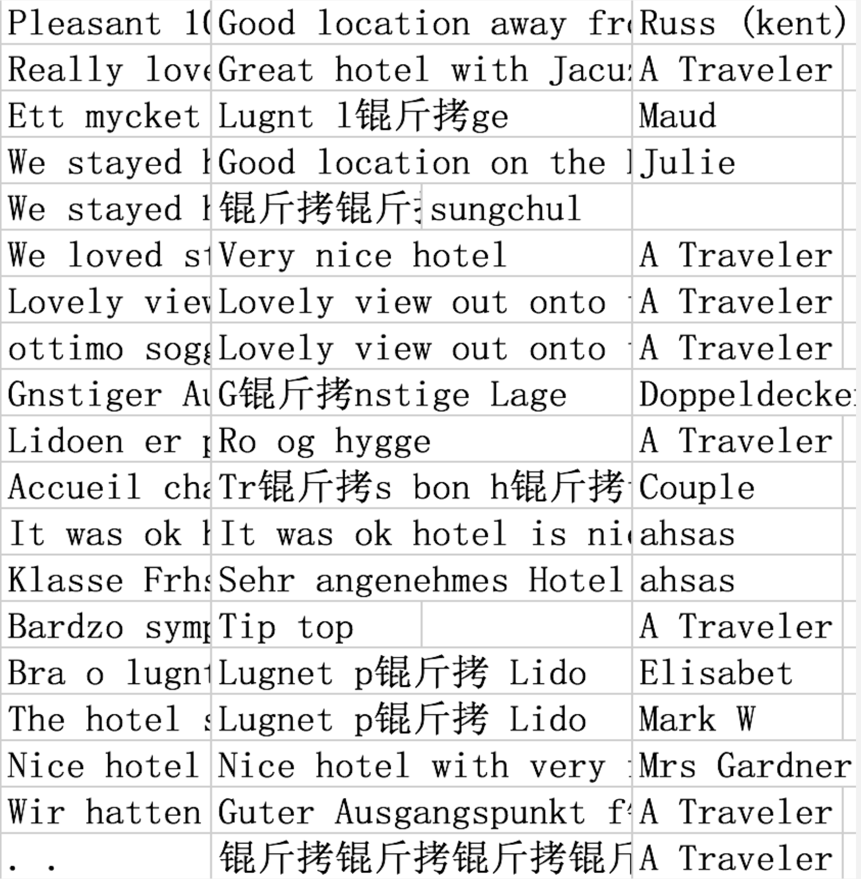
How

Use nltk from python to do text mining of text review of customers.

PART 02

Data Preprocessing

Messy code



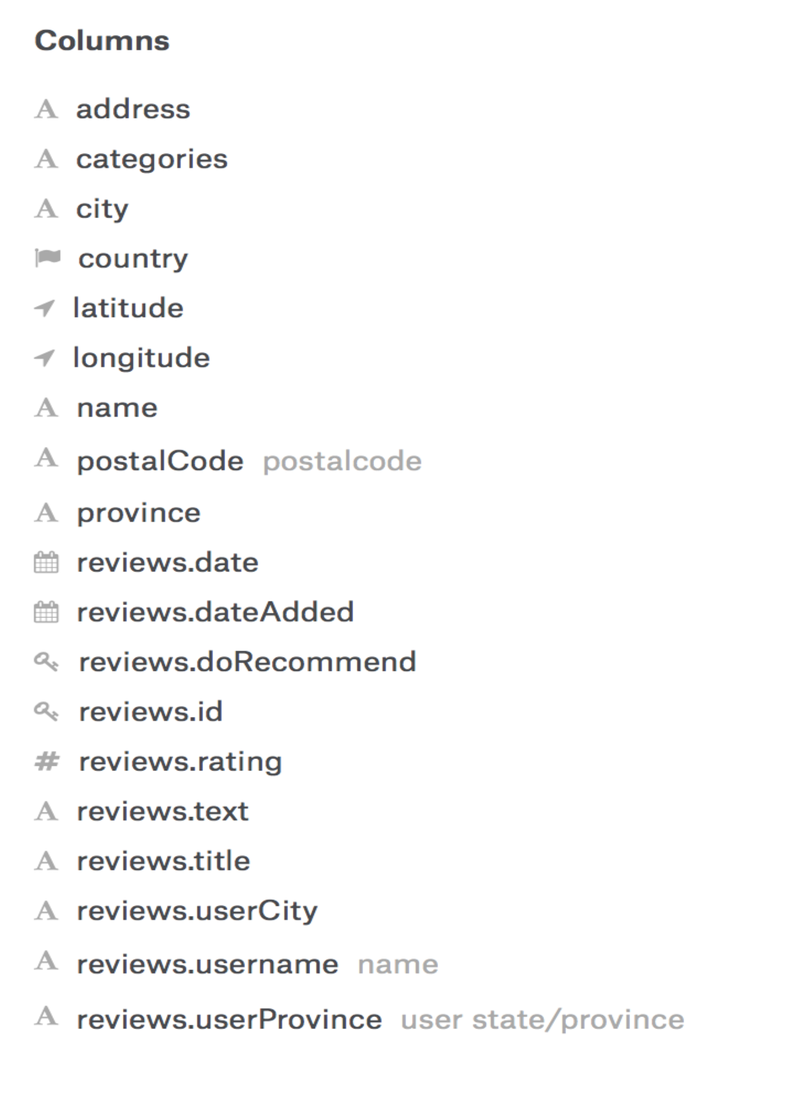
Pleasant 1(	Good location away fr	Russ (kent)
Really love	Great hotel with Jacu	A Traveler
Ett mycket	Lugnt l锟斤拷ge	Maud
We stayed h	Good location on the	Julie
We stayed h	锟斤拷锟斤 sungchul	
We loved st	Very nice hotel	A Traveler
Lovely view	Lovely view out onto	A Traveler
ottimo sogg	Lovely view out onto	A Traveler
Gnstiger Au	G锟斤拷nstige Lage	Doppeldecke
Lidoen er p	Ro og hygge	A Traveler
Accueil cha	Tr锟斤拷s bon h锟斤拷	Couple
It was ok h	It was ok hotel is ni	ahsas
Klasse Frhs	Sehr angenehmes Hotel	ahsas
Bardzo symp	Tip top	A Traveler
Bra o lugnt	Lugnet p锟斤拷 Lido	Elisabet
The hotel s	Lugnet p锟斤拷 Lido	Mark W
Nice hotel	Nice hotel with very	Mrs Gardner
Wir hatten	Guter Ausgangspunkt f	A Traveler
. .	锟斤拷锟斤拷锟斤拷锟斤	A Traveler

- Drop the blank values
- Use the regex to filter the messy code

Data Preprocessing and Cleaning

- Remove some columns

- Use nltk package to lowercase, and remove the stop words and punctuation of text document

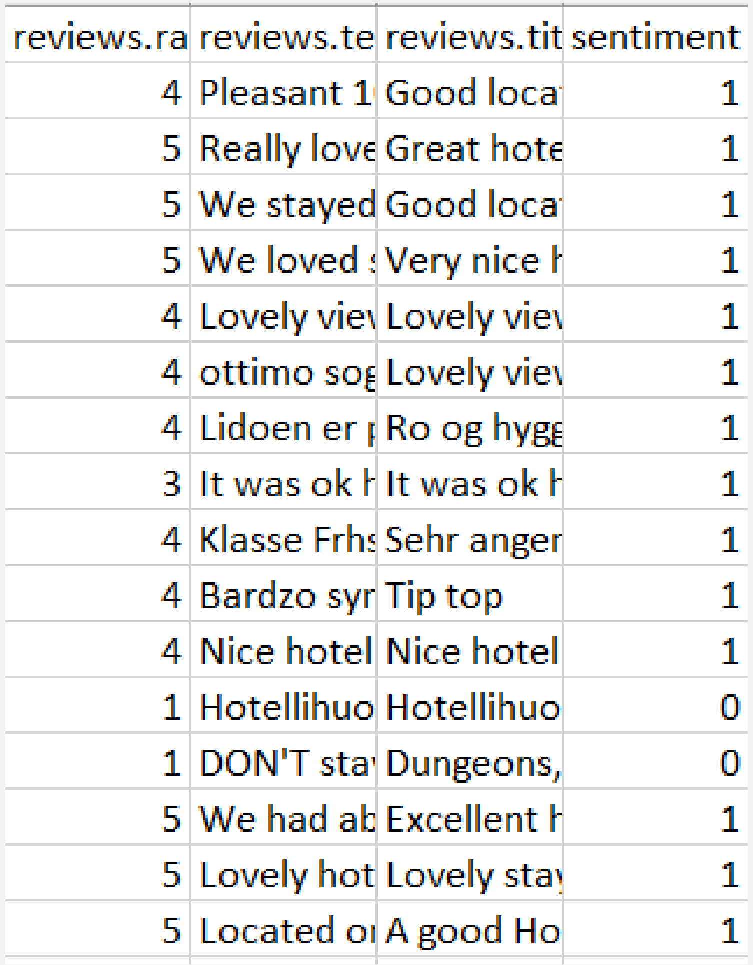


Columns

A address
A categories
A city
country
latitude
longitude
A name
A postalCode postalcode
A province
reviews.date
reviews.dateAdded
reviews.doRecommend
reviews.id
# reviews.rating
A reviews.text
A reviews.title
A reviews.userCity
A reviews.username name
A reviews.userProvince user state/province

Set Sentiment Feature

- Rating greater than 3 → good → 1
- Rating smaller than or equal to 3 → bad → 0
- Unbalanced data: 1 → 21406; 0 → 4567



reviews.ra	reviews.te	reviews.tit	sentiment
4	Pleasant 1	Good loca	1
5	Really love	Great hote	1
5	We stayed	Good loca	1
5	We loved s	Very nice h	1
4	Lovely viev	Lovely viev	1
4	ottimo sog	Lovely viev	1
4	Lidoen er p	Ro og hygg	1
3	It was ok h	It was ok h	1
4	Klasse Frhs	Sehr anger	1
4	Bardzo syr	Tip top	1
4	Nice hotel	Nice hotel	1
1	Hotellihuo	Hotellihuo	0
1	DON'T sta	Dungeons,	0
5	We had ab	Excellent h	1
5	Lovely hot	Lovely stay	1
5	Located or	A good Ho	1

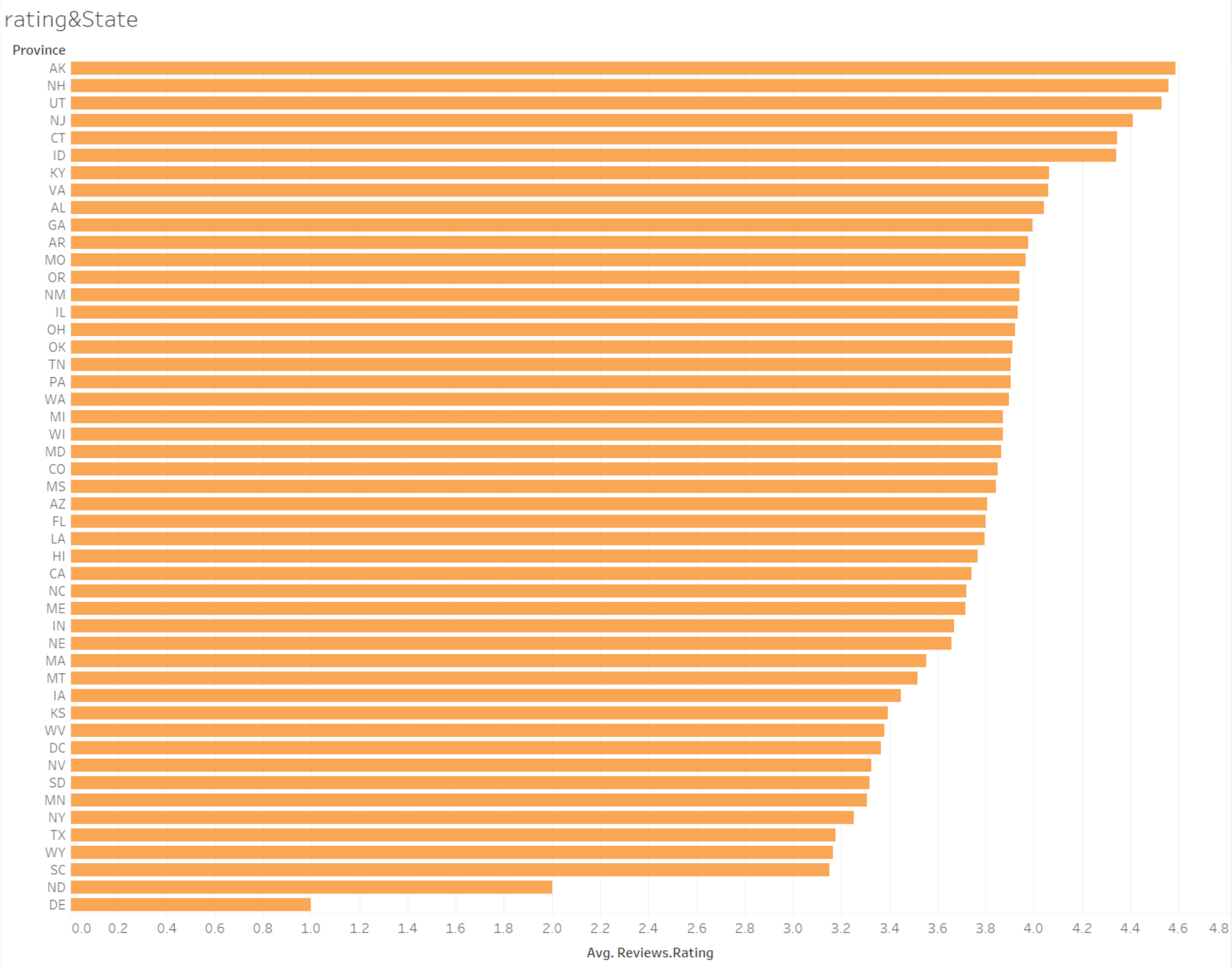
PART 03

Visualization & Analytics

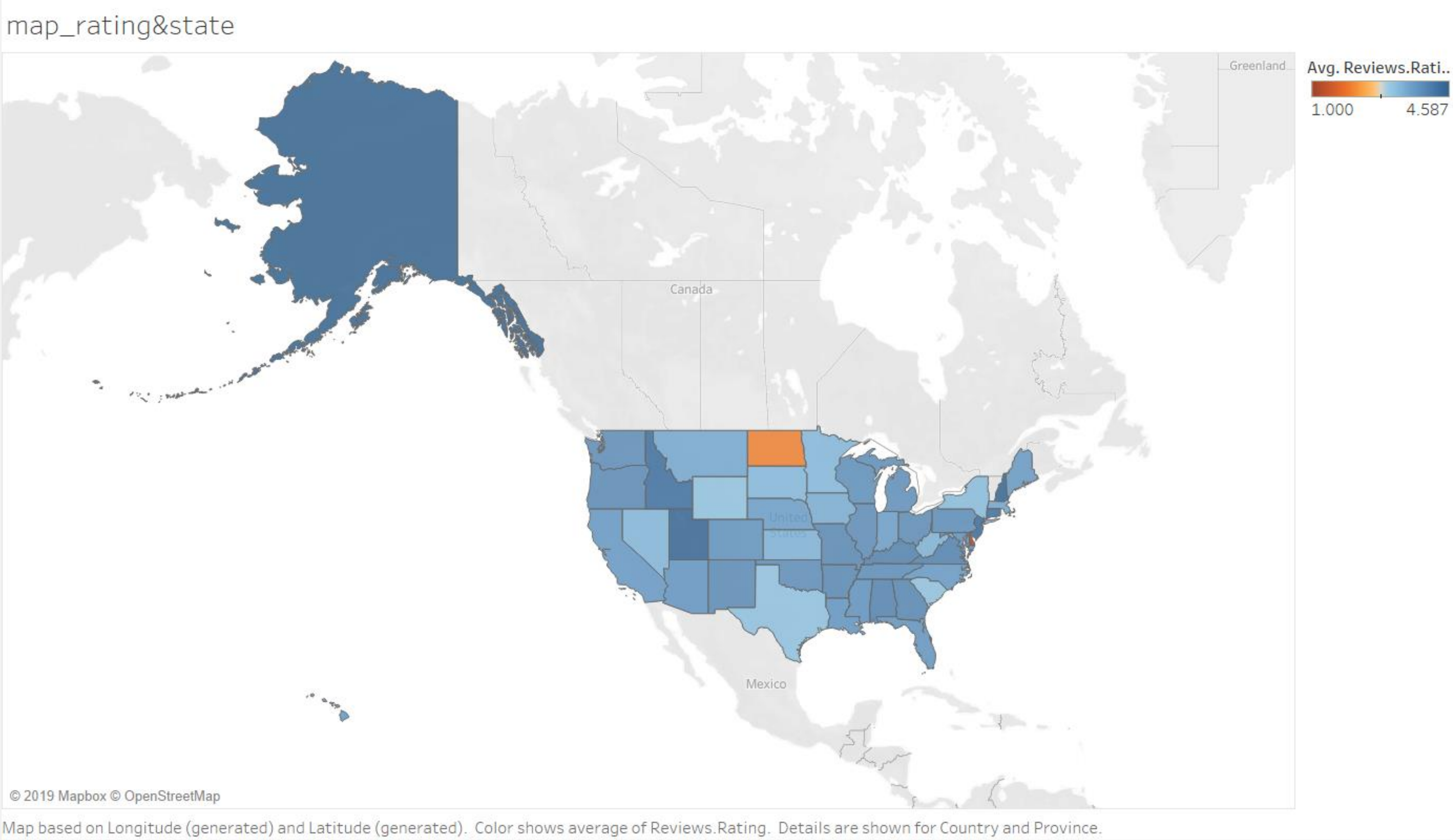
Rating by state

TOP 5:

1. Alaska
2. New Hampshire
3. Utah
4. New Jersey
5. Connecticut

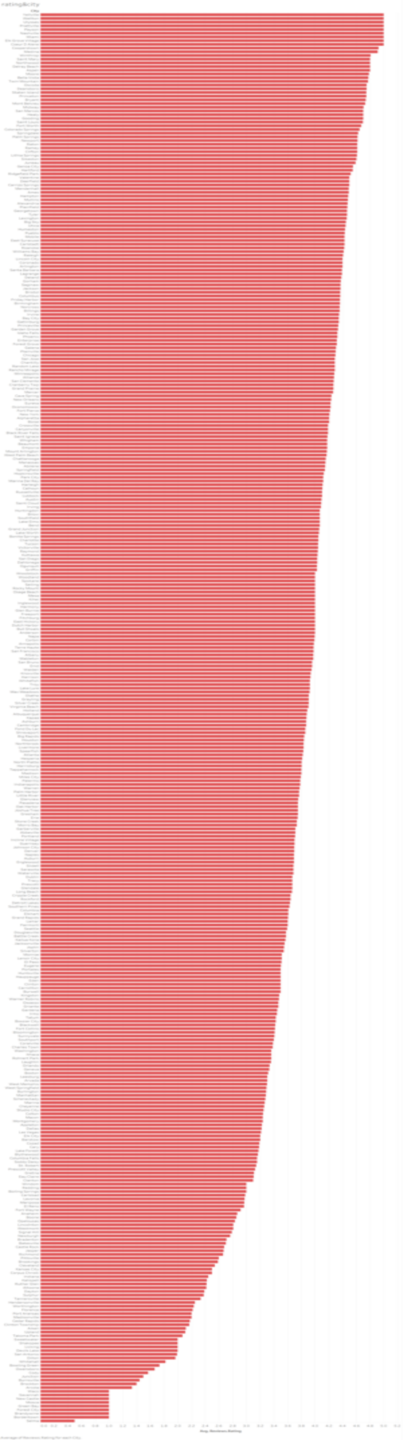


Average of Reviews.Rating for each Province.

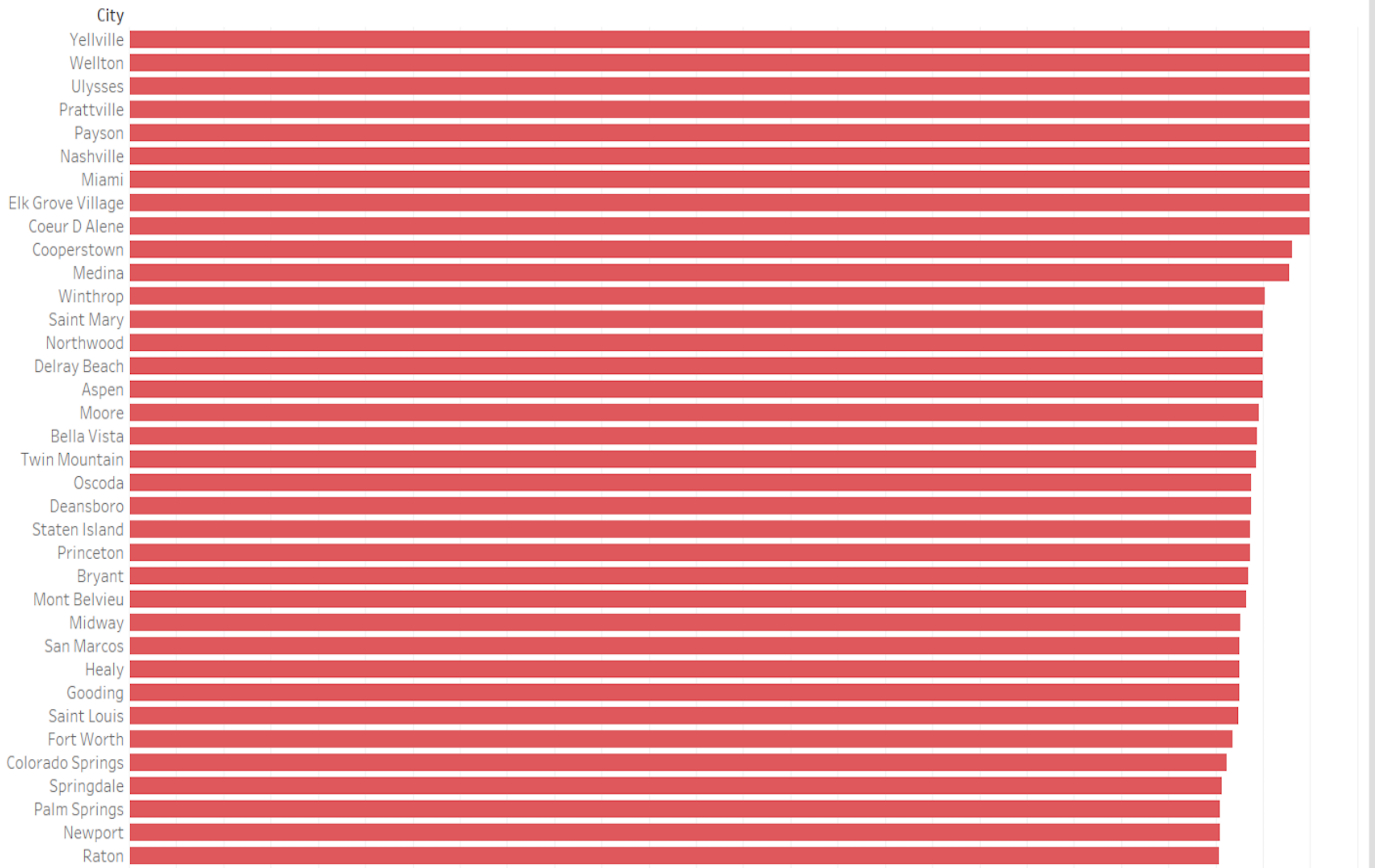


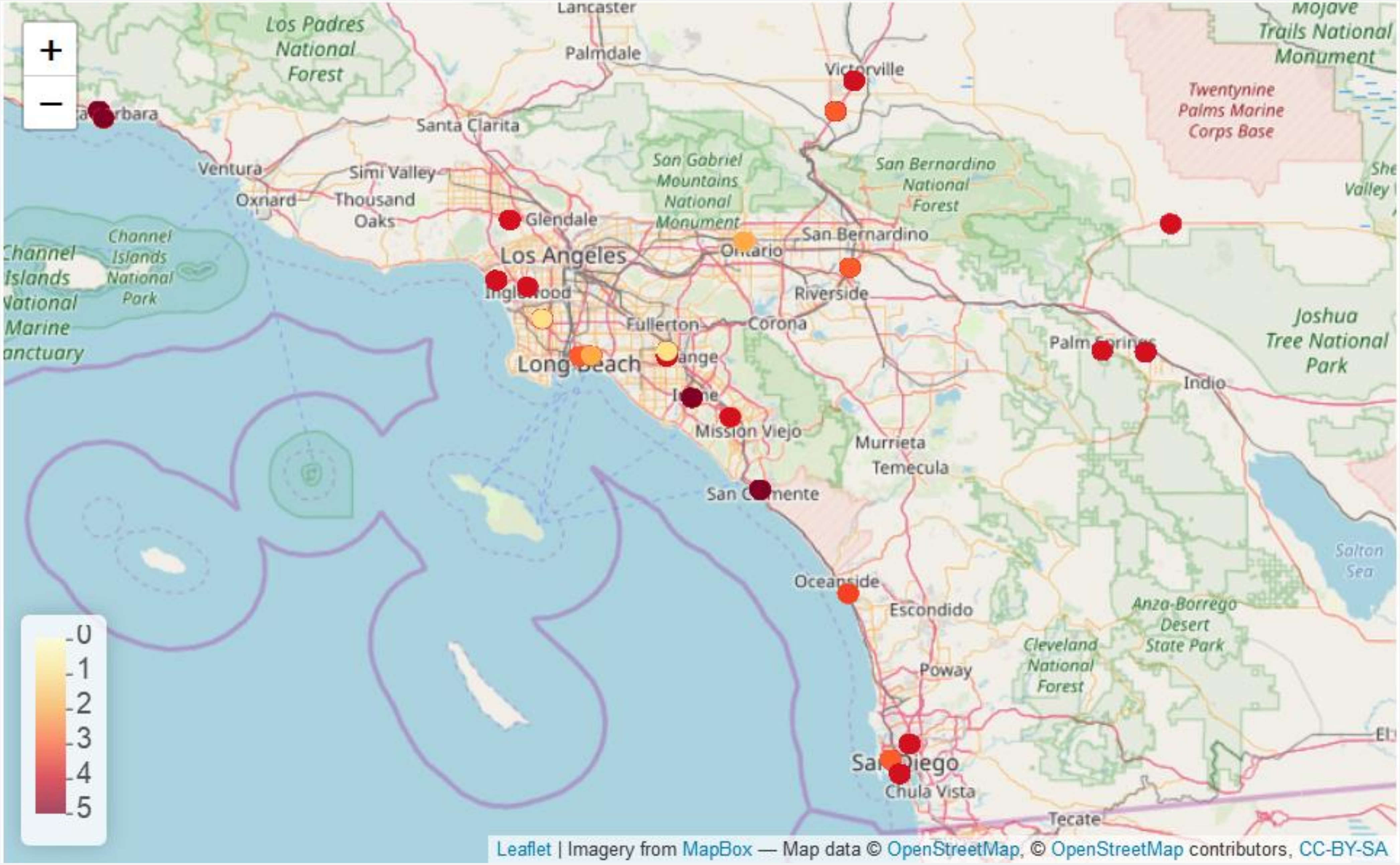
map_rating&state

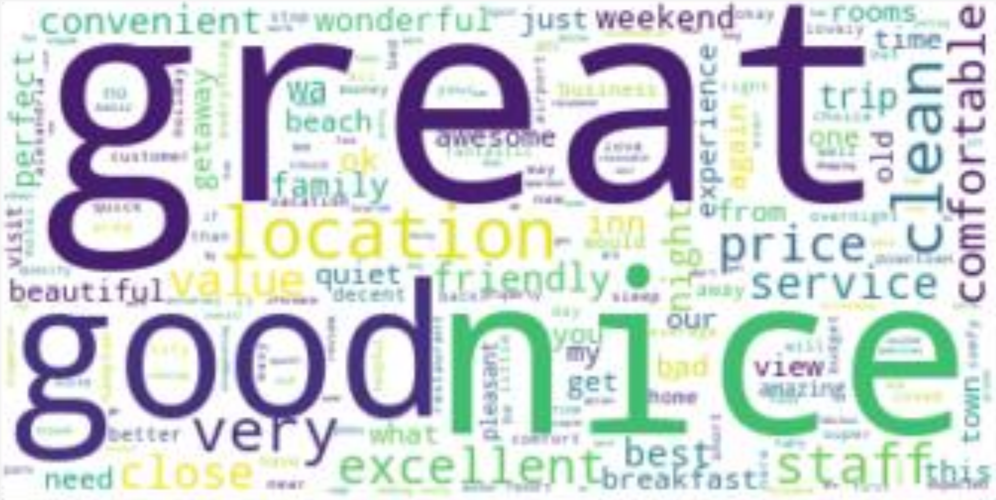
Map based on Longitude (generated) and Latitude (generated). Color shows average of Reviews.Rating. Details are shown for Country and Province.

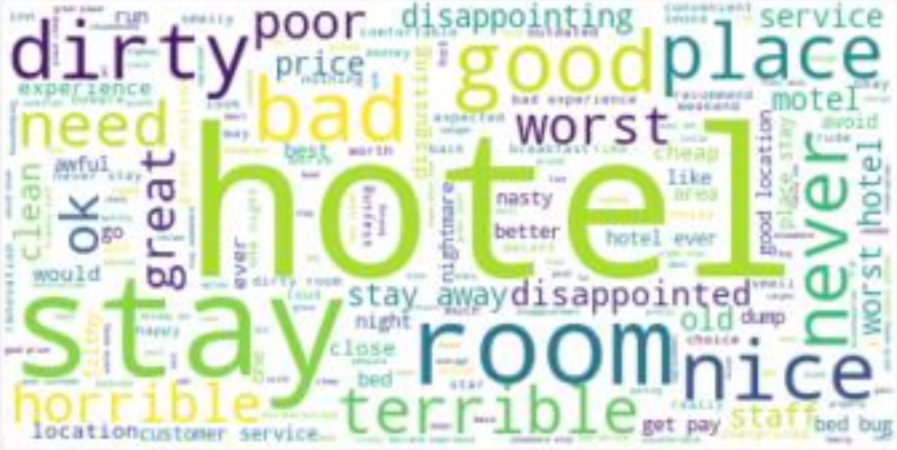


rating&city

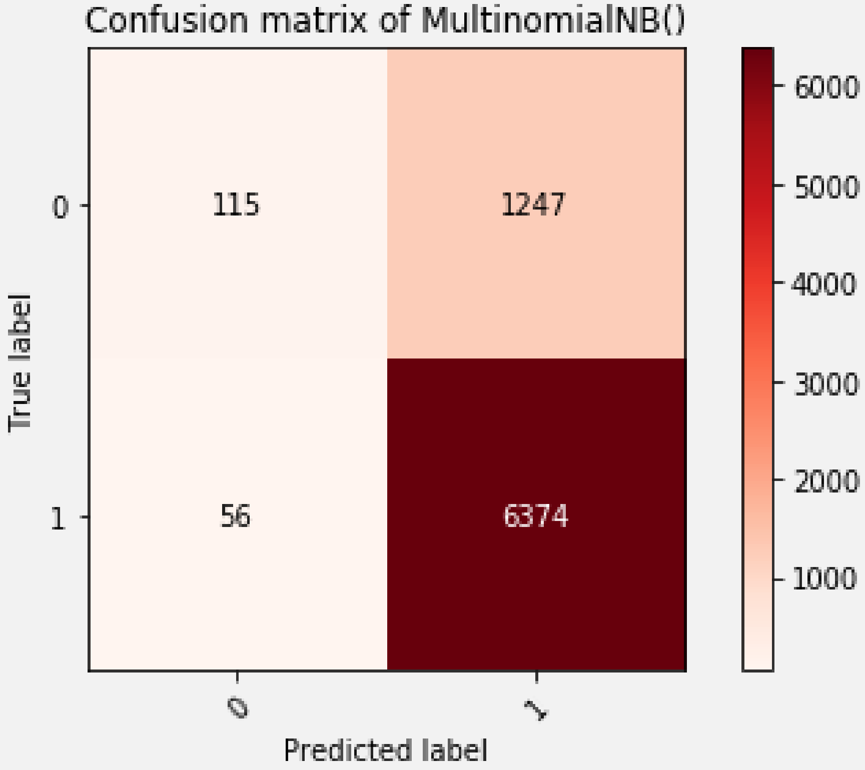

City
Yellville
Wellton
Ulysses
Prattville
Payson
Nashville
Miami
Elk Grove Village
Coeur D Alene
Cooperstown
Medina
Winthrop
Saint Mary
Northwood
Delray Beach
Aspen
Moore
Bella Vista
Twin Mountain
Oscoda
Deansboro
Staten Island
Princeton
Bryant
Mont Belvieu
Midway
San Marcos
Healy
Gooding
Saint Louis
Fort Worth
Colorado Springs
Springdale
Palm Springs
Newport
Raton

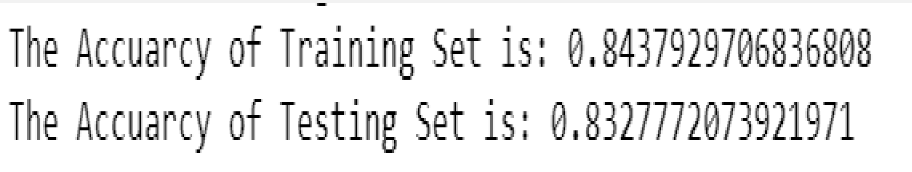

+
−
0
1
2
3
4
5
Lancaster
Palmdale
Los Padres National Forest
Santa Barbara
Santa Clarita
Ventura
Oxnard
Simi Valley
Thousand Oaks
Glendale
Los Angeles
Inglewood
Long Beach
Fullerton
Orange
Irvine
Mission Viejo
San Clemente
Corona
Riverside
Ontario
San Bernardino
San Bernardino National Forest
San Gabriel Mountains National Monument
Victorville
Murrieta
Temecula
Oceanside
Escondido
Poway
San Diego
Chula Vista
Tecate
Palm Springs
Indio
Joshua Tree National Park
Mojave Trails National Monument
Twentynine Palms Marine Corps Base
Salton Sea
Anza-Borrego Desert State Park
Cleveland National Forest
Channel Islands National Park
Channel Islands National Marine Sanctuary
Leaflet | Imagery from MapBox — Map data © OpenStreetMap, © OpenStreetMap contributors, CC-BY-SA


great
good
nice
location
clean
comfortable
price
service
friendly
value
staff
excellent
close
very
best
breakfast
convenient
wonderful
just
weekend
rooms
time
trip
perfect
getaway
beach
family
awesome
experience
beautiful
quiet
decent
view
amazing
pleasant
town
need
inn
night
bad
home
our
get
you
my
what
business
ok
wa
one
old
again
from
visit
this

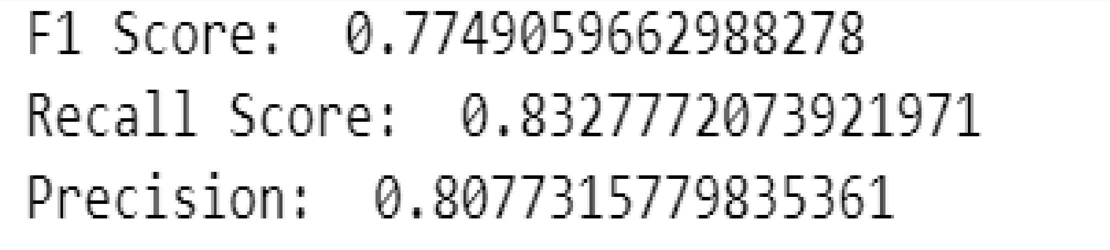

hotel
stay
room
nice
dirty
poor
disappointing
service
good
place
bad
worst
need
motel
never
great
ok
clean
terrible
horrible
staff
price
experience
awful
nasty
better
hotel ever
worst hotel
stay away
disappointed
old
dump
close
night
bed
location
customer service
get pay
bed bug
cheap
like
area
nightmare
disgusting
would
go
good location

Naïve Bayes-Multinomial Naive Bayes

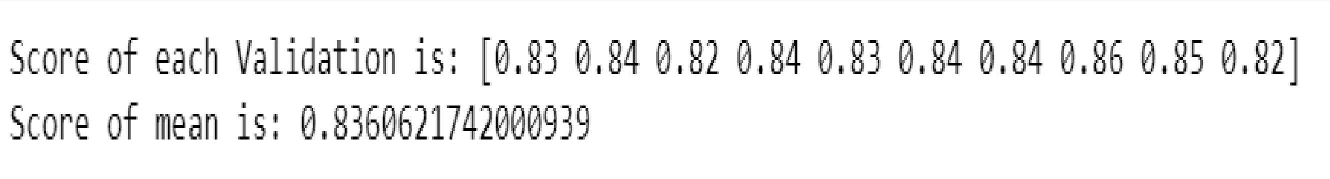




```
The Accuarcy of Training Set is: 0.8437929706836808
The Accuarcy of Testing Set is: 0.8327772073921971
```

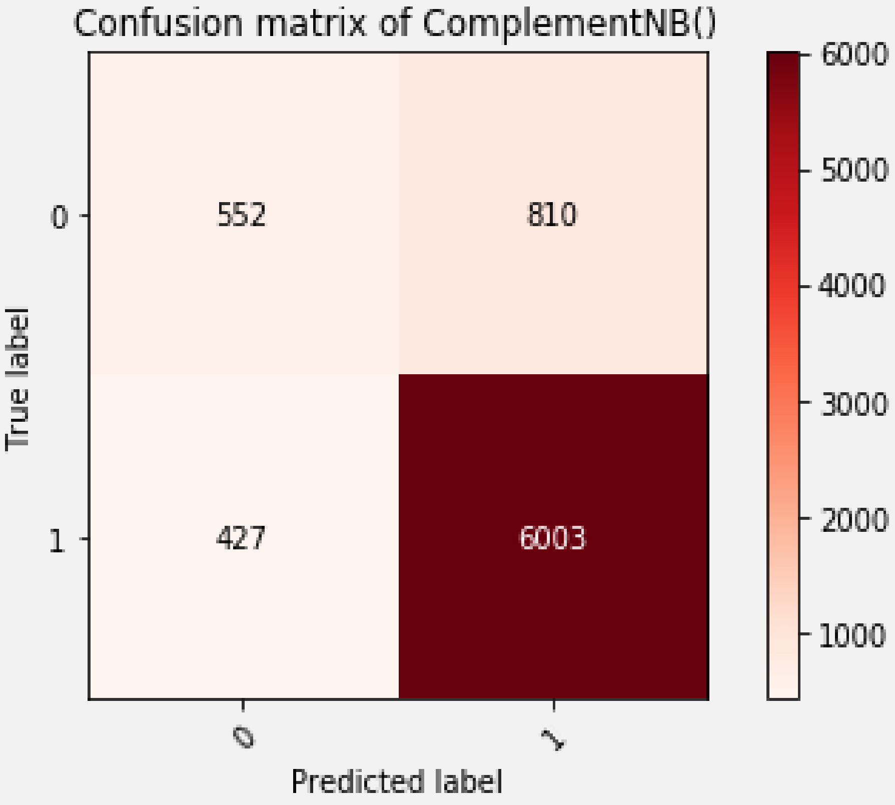


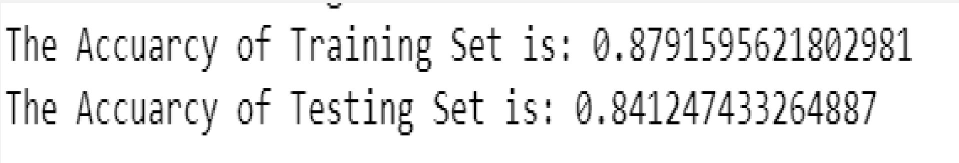
```
F1 Score:  0.7749059662988278
Recall Score:  0.8327772073921971
Precision:  0.8077315779835361
```



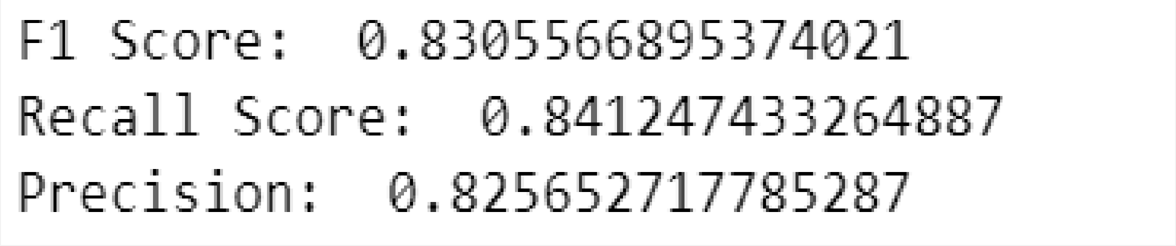
```
Score of each Validation is: [0.83 0.84 0.82 0.84 0.83 0.84 0.84 0.86 0.85 0.82]
Score of mean is: 0.8360621742000939
```

Naïve Bayes-Complement Naive Bayes

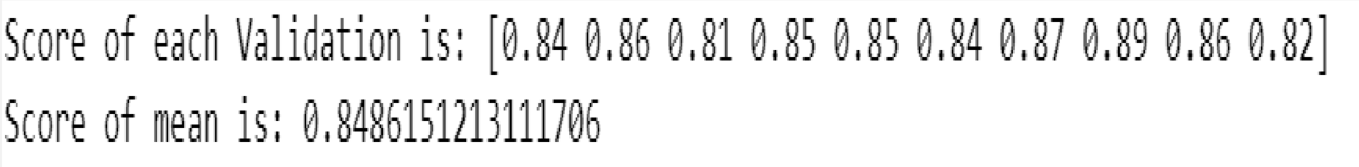




```
The Accuarcy of Training Set is: 0.8791595621802981
The Accuarcy of Testing Set is: 0.841247433264887
```

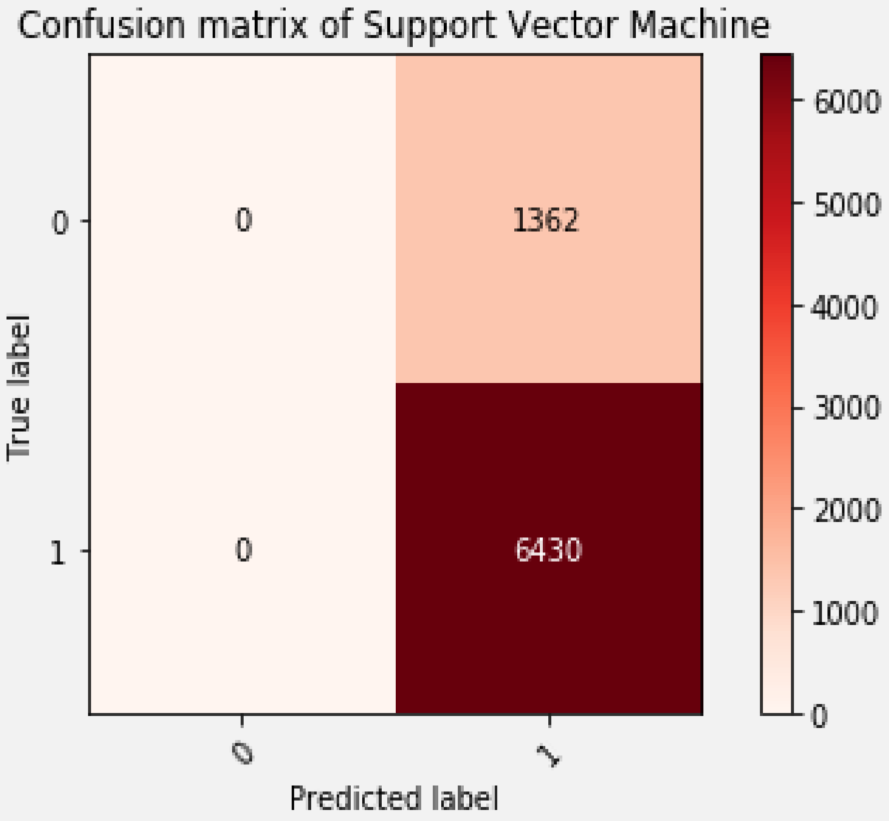


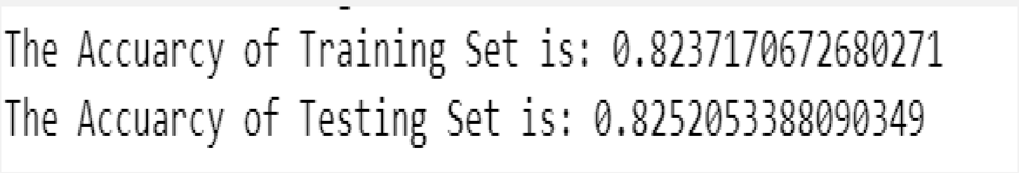
```
F1 Score:  0.8305566895374021
Recall Score:  0.841247433264887
Precision:  0.825652717785287
```



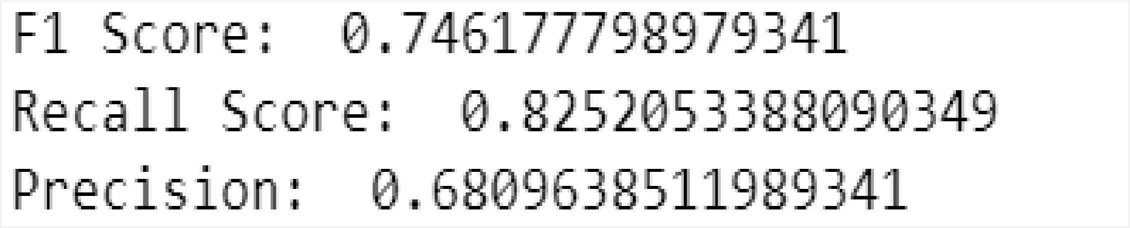
```
Score of each Validation is: [0.84 0.86 0.81 0.85 0.85 0.84 0.87 0.89 0.86 0.82]
Score of mean is: 0.8486151213111706
```

Support Vector Machine

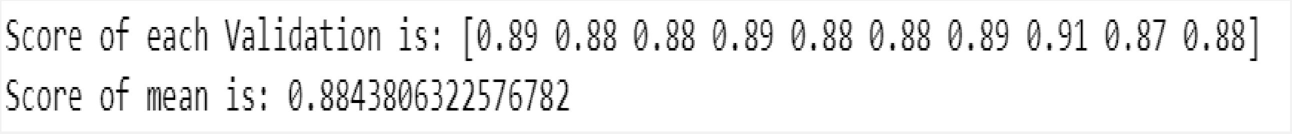




```
The Accuarcy of Training Set is: 0.8237170672680271
The Accuarcy of Testing Set is: 0.8252053388090349
```

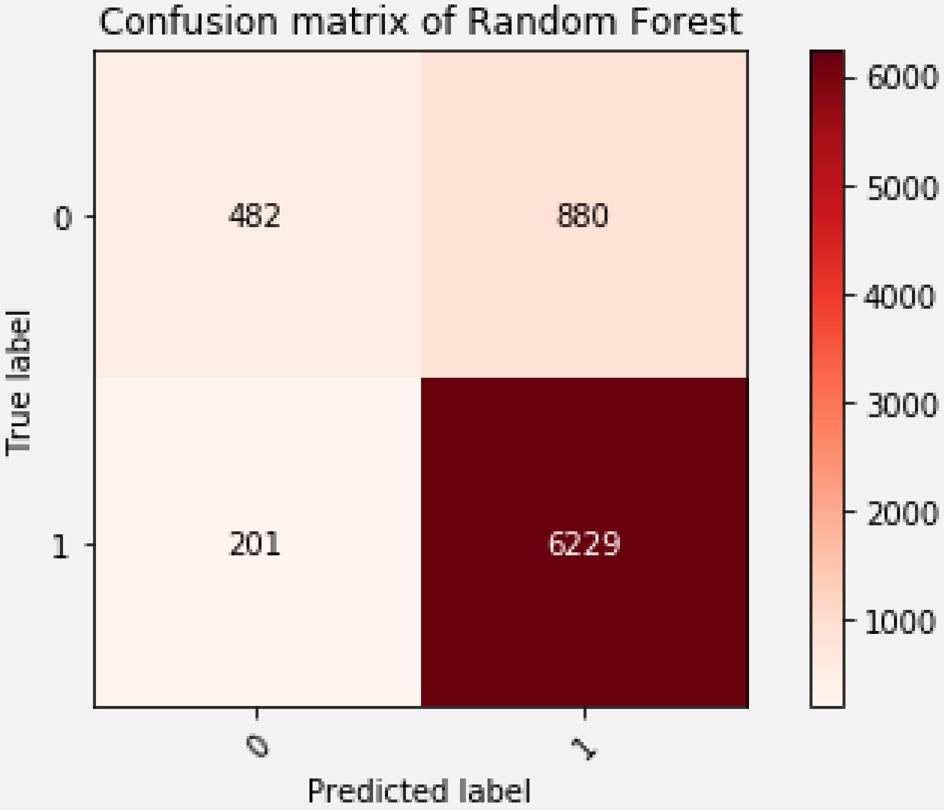


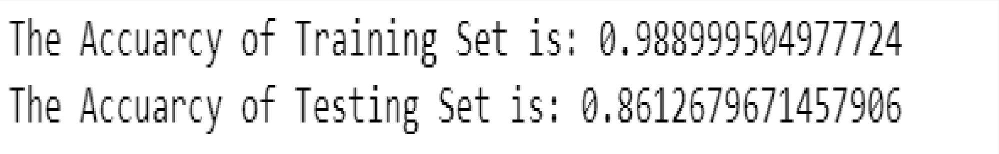
```
F1 Score:  0.746177798979341
Recall Score:  0.8252053388090349
Precision:  0.6809638511989341
```



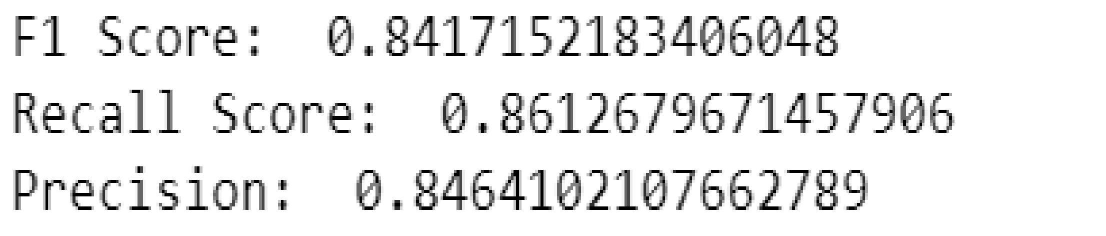
```
Score of each Validation is: [0.89 0.88 0.88 0.89 0.88 0.88 0.89 0.91 0.87 0.88]
Score of mean is: 0.8843806322576782
```

Random Forest

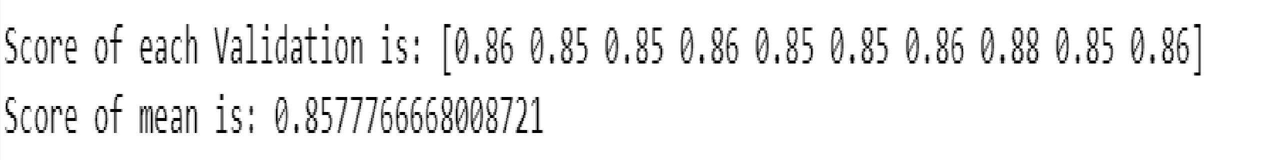




```
The Accuarcy of Training Set is: 0.988999504977724
The Accuarcy of Testing Set is: 0.8612679671457906
```

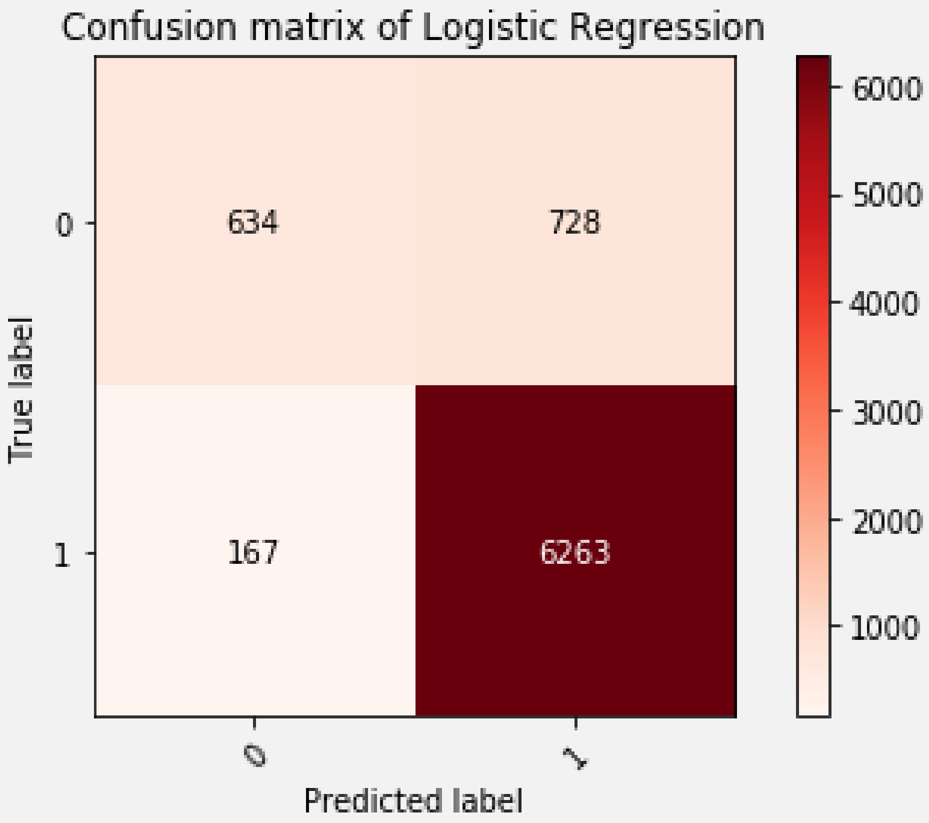


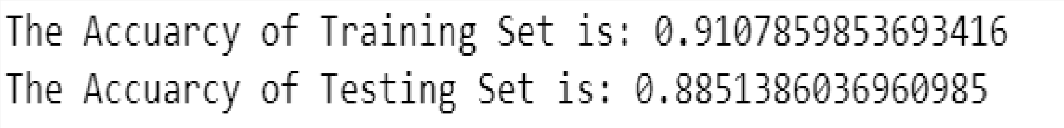
```
F1 Score:  0.8417152183406048
Recall Score:  0.8612679671457906
Precision:  0.8464102107662789
```



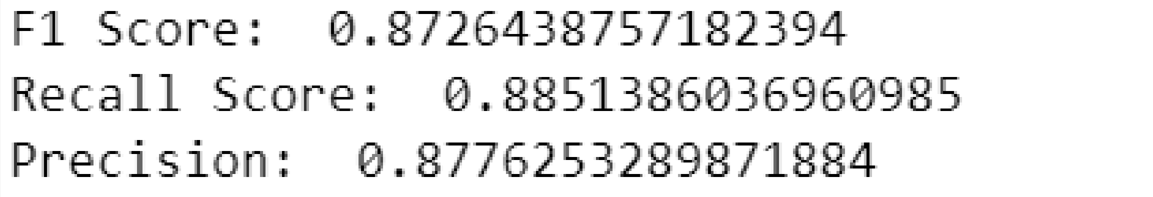
```
Score of each Validation is: [0.86 0.85 0.85 0.86 0.85 0.85 0.86 0.88 0.85 0.86]
Score of mean is: 0.8577766668008721
```

Logistic Regression

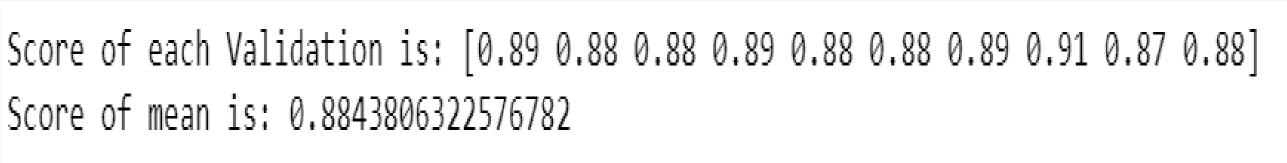




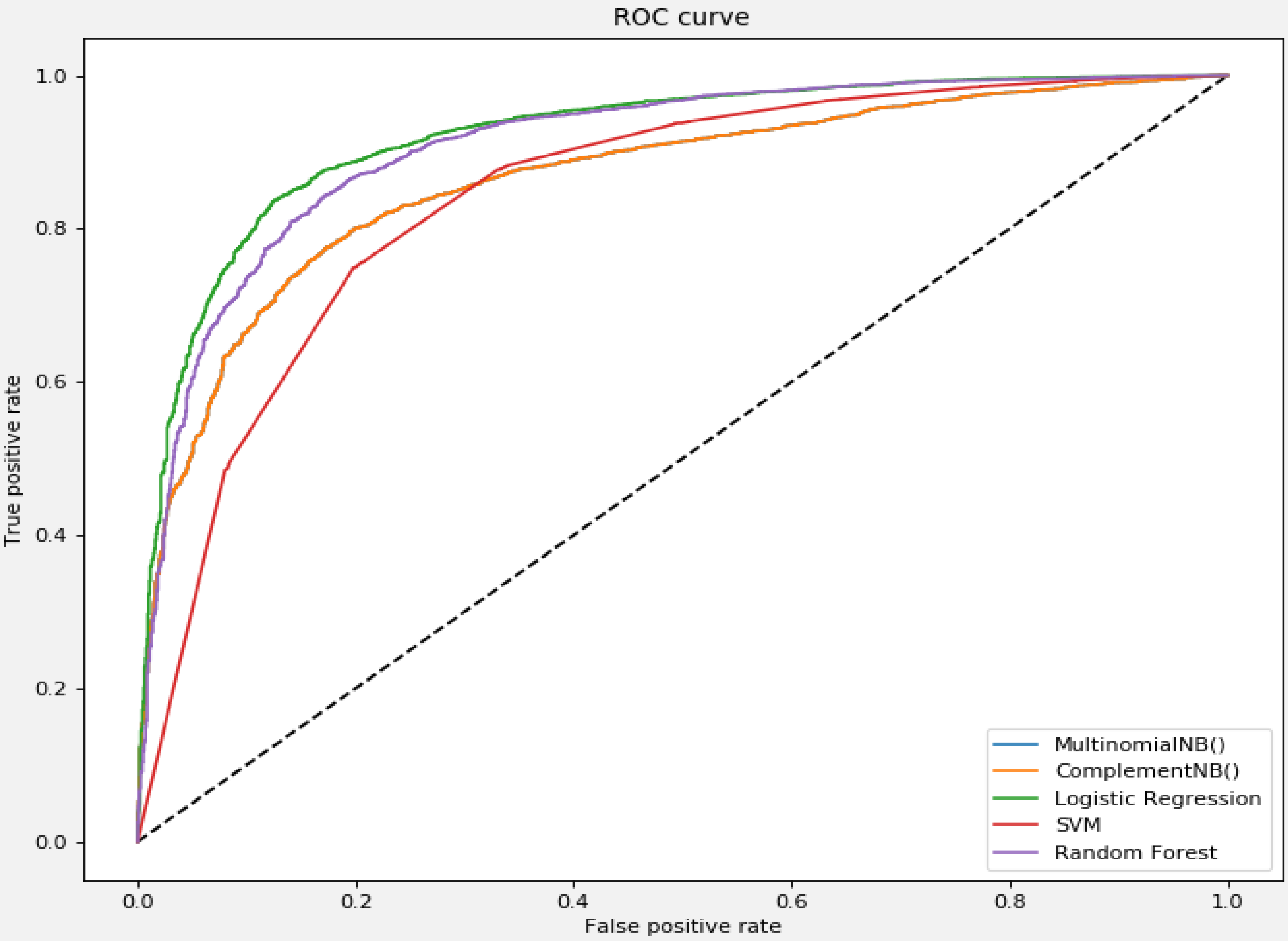
```
The Accuarcy of Training Set is: 0.9107859853693416
The Accuarcy of Testing Set is: 0.8851386036960985
```



```
F1 Score:  0.8726438757182394
Recall Score:  0.8851386036960985
Precision:  0.8776253289871884
```



```
Score of each Validation is: [0.89 0.88 0.88 0.89 0.88 0.88 0.89 0.91 0.87 0.88]
Score of mean is: 0.8843806322576782
```


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
MultinomialNB()
ComplementNB()
Logistic Regression
SVM
Random Forest

Overview of Algorithms

	MultinomialNB()	ComplementNB()	Logistic Regression	SVM	Random Forest
Training Set	0.8437	0.8791	0.9107	0.8237	0.9889
Test Set	0.8327	0.8412	0.8851	0.8252	0.8612
Mean of Cross Validation	0.8360	0.8486	0.8843	0.8843	0.8577
F1 Score	0.7749	0.8305	0.8726	0.7461	0.8417
Recall Score	0.8327	0.8412	0.8851	0.8252	0.8612
Precision	0.8077	0.8256	0.8776	0.6809	0.8464
Training Time	0.0362’s	0.0195’s	1.3972’s	421.3718’s	14.5134’s

PART 04

Future Work

Good experience

Good recommendation

Best hotel

Thank You