Hotel Reviews Analysis

Team Pandas Learning

Mary Kiangi

Hoa Duong

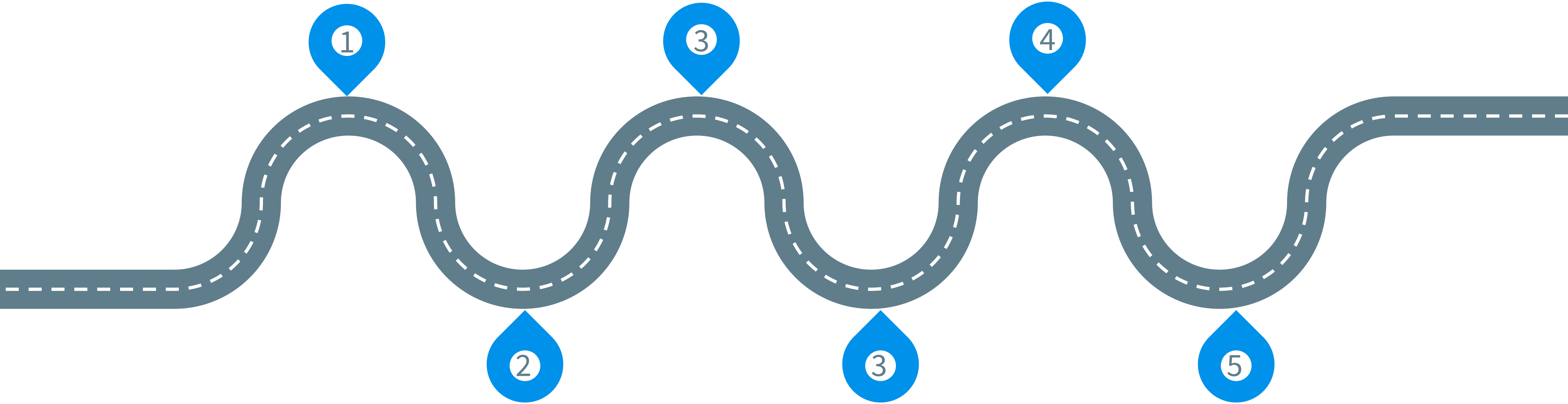
Joseph High

Jiawen Li

Ziyun Lu

https://www.shutterstock.com/image-vector/funny-panda-reads-book-style-set-2030376248
https://www.vectorstock.com/royalty-free-vector/cute-panda-reading-a-book-vector-16631904
https://www.vectorstock.com/royalty-free-vector/panda-reading-a-book-vector-16810734

Outline



Introduction

Data Cleaning

Exploratory Data Analysis by Rating

Exploratory Data Analysis by Reviews

Predictive Models

Future Work

1. Introduction

The Traveler’s Question

“I want to travel to Los Angeles. What should I expect out of the hotels there?”

The Owner’s Question

“What do my customers care about in my property?”

Datasets

Hotel reviews

◎ 55,912 observations

Hotels information

- Name
- Category
- Address
- Location
- Coordinates

Reviews information

- Date
- Numerical rating
- Title and text of reviews
- Reviewer’s name
- Reviewer’s location

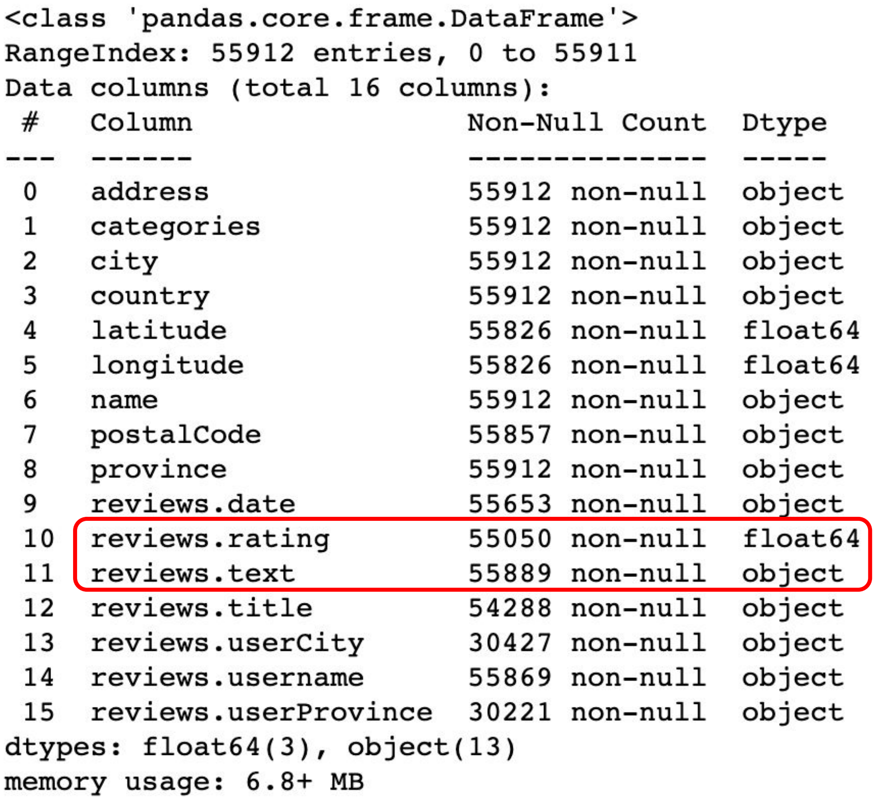
Datasets

Review example

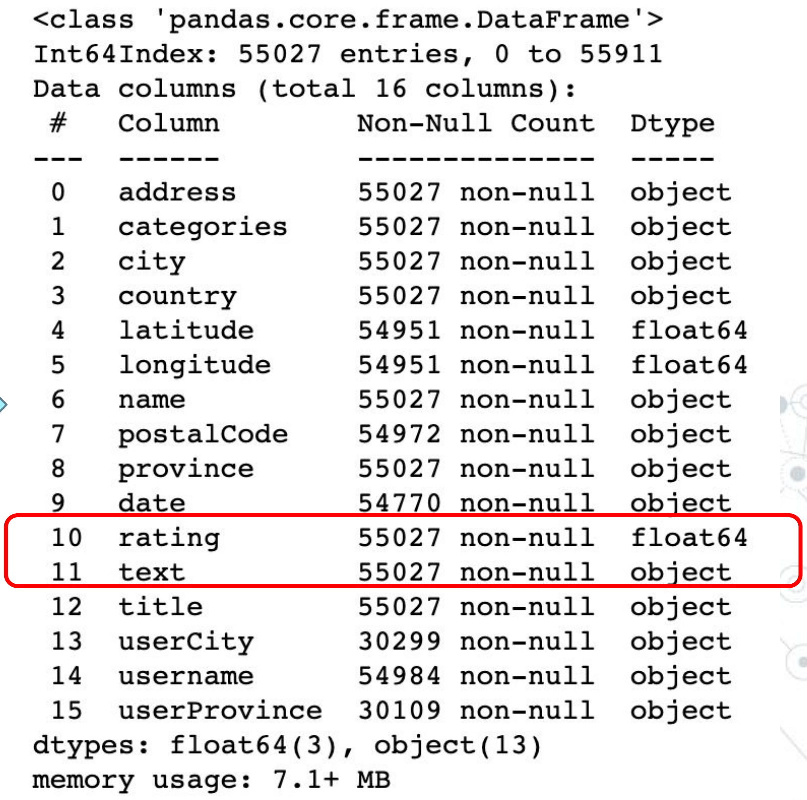
Review	Rating
“We stayed here for four nights in October. The hotel staff were welcoming, friendly and helpful. Assisted in booking tickets for the opera. The rooms were clean and comfortable- good shower, light and airy rooms with windows you could open wide. Beds were comfortable. Plenty of choice for breakfast. Spa at hotel nearby which we used while we were there.”	5.0
“Took more than 2 hour waiting to check-in to our room. Otherwise stay was comfortable. Have issue with slow Wi-Fi connection.”	3.0
“Sheets were filthy, jacuzzi was freezing with bugs in it, the breakfast had options but the eggs and bacon tasted terrible!”	1.0

2. Data Cleaning

Initial Data Cleaning

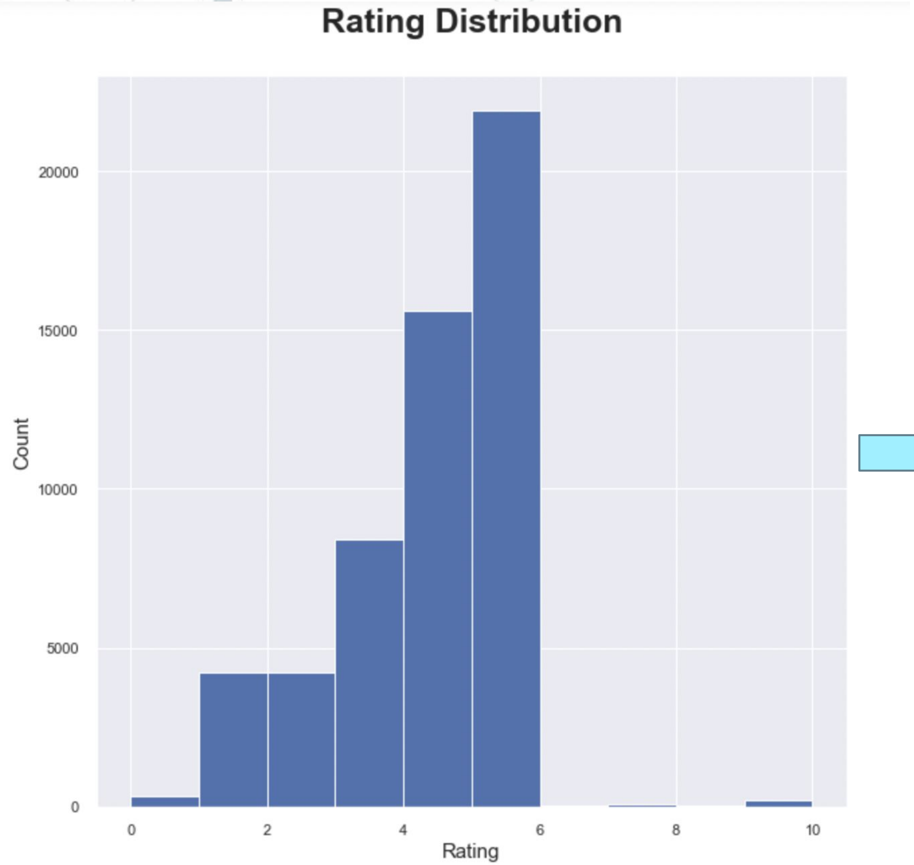


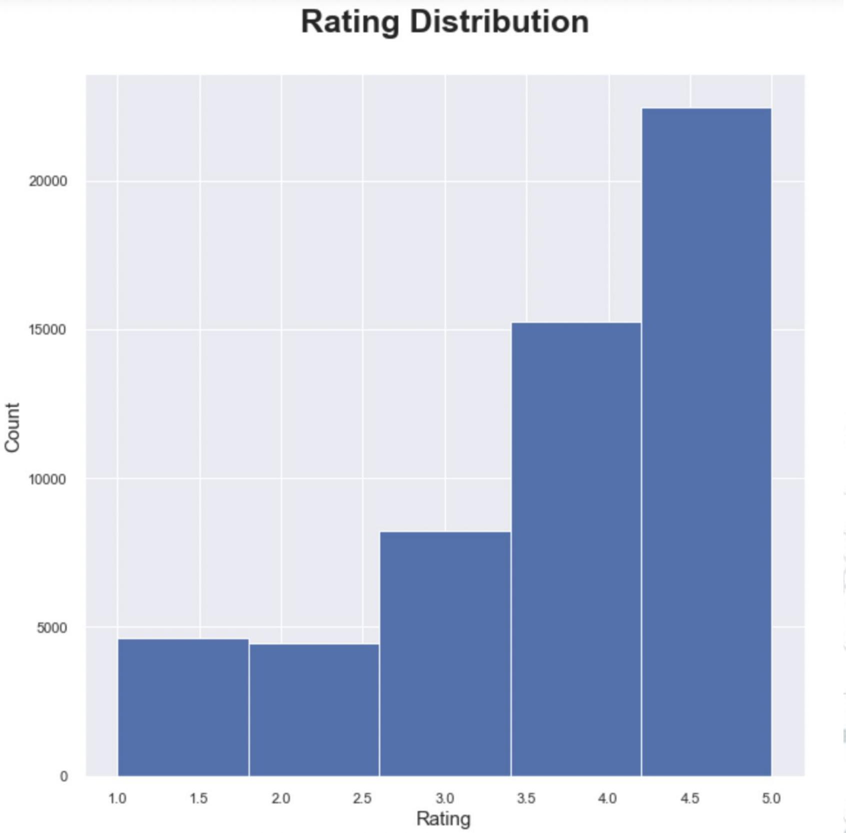
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55912 entries, 0 to 55911
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   address               55912 non-null  object
 1   categories            55912 non-null  object
 2   city                  55912 non-null  object
 3   country               55912 non-null  object
 4   latitude              55826 non-null  float64
 5   longitude             55826 non-null  float64
 6   name                  55912 non-null  object
 7   postalCode            55857 non-null  object
 8   province              55912 non-null  object
 9   reviews.date          55653 non-null  object
 10  reviews.rating        55050 non-null  float64
 11  reviews.text          55889 non-null  object
 12  reviews.title         54288 non-null  object
 13  reviews.userCity      30427 non-null  object
 14  reviews.username      55869 non-null  object
 15  reviews.userProvince  30221 non-null  object
dtypes: float64(3), object(13)
memory usage: 6.8+ MB
```



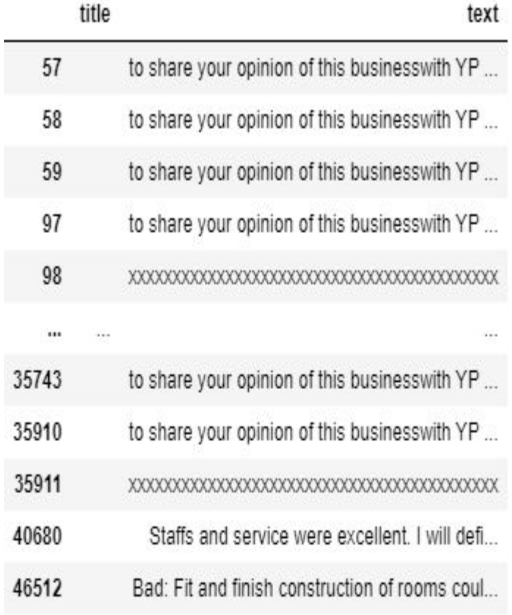
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 55027 entries, 0 to 55911
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   address        55027 non-null  object
 1   categories     55027 non-null  object
 2   city           55027 non-null  object
 3   country        55027 non-null  object
 4   latitude       54951 non-null  float64
 5   longitude      54951 non-null  float64
 6   name           55027 non-null  object
 7   postalCode     54972 non-null  object
 8   province       55027 non-null  object
 9   date           54770 non-null  object
 10  rating         55027 non-null  float64
 11  text           55027 non-null  object
 12  title          55027 non-null  object
 13  userCity       30299 non-null  object
 14  username       54984 non-null  object
 15  userProvince   30109 non-null  object
dtypes: float64(3), object(13)
memory usage: 7.1+ MB
```

Initial Data Cleaning

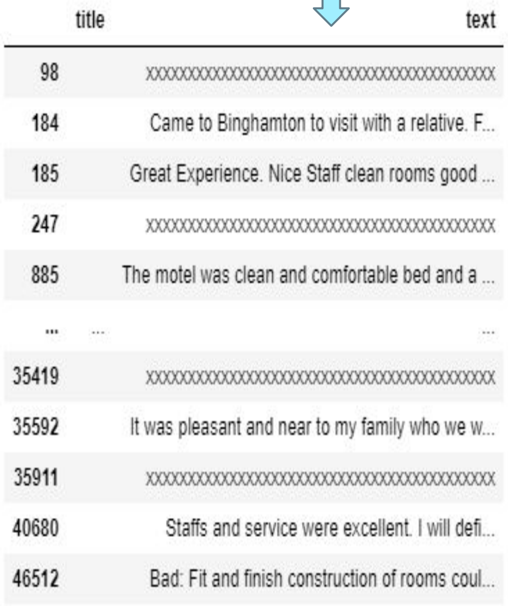




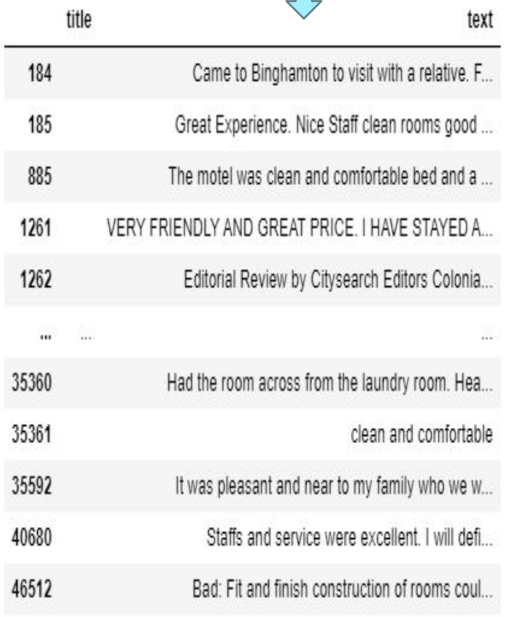
Data Cleaning - Reviews



	title	text
57		to share your opinion of this businesswith YP ...
58		to share your opinion of this businesswith YP ...
59		to share your opinion of this businesswith YP ...
97		to share your opinion of this businesswith YP ...
98		xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
...	...	...
35743		to share your opinion of this businesswith YP ...
35910		to share your opinion of this businesswith YP ...
35911		xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
40680		Staffs and service were excellent. I will defi...
46512		Bad: Fit and finish construction of rooms coul...

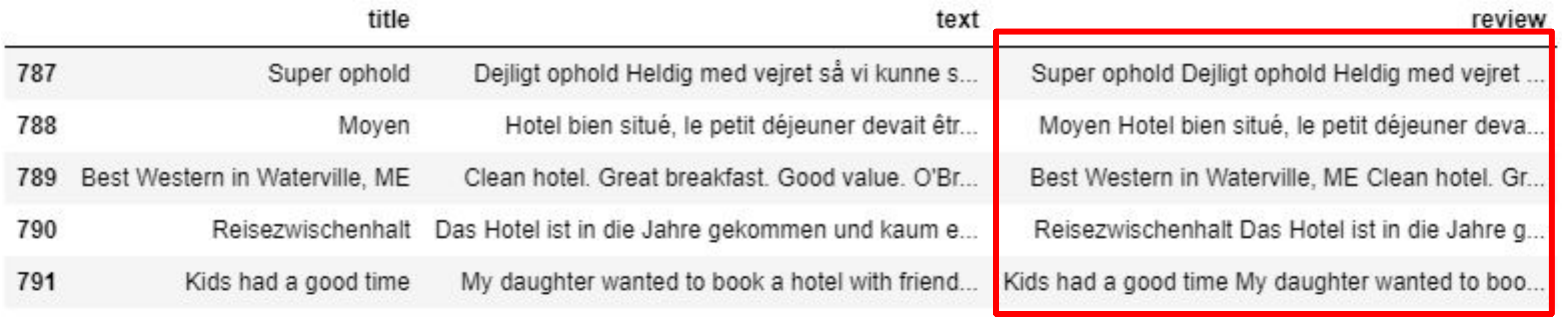


	title	text
98		xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
184		Came to Binghamton to visit with a relative. F...
185		Great Experience. Nice Staff clean rooms good ...
247		xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
885		The motel was clean and comfortable bed and a ...
...	...	...
35419		xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
35592		It was pleasant and near to my family who we w...
35911		xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
40680		Staffs and service were excellent. I will defi...
46512		Bad: Fit and finish construction of rooms coul...

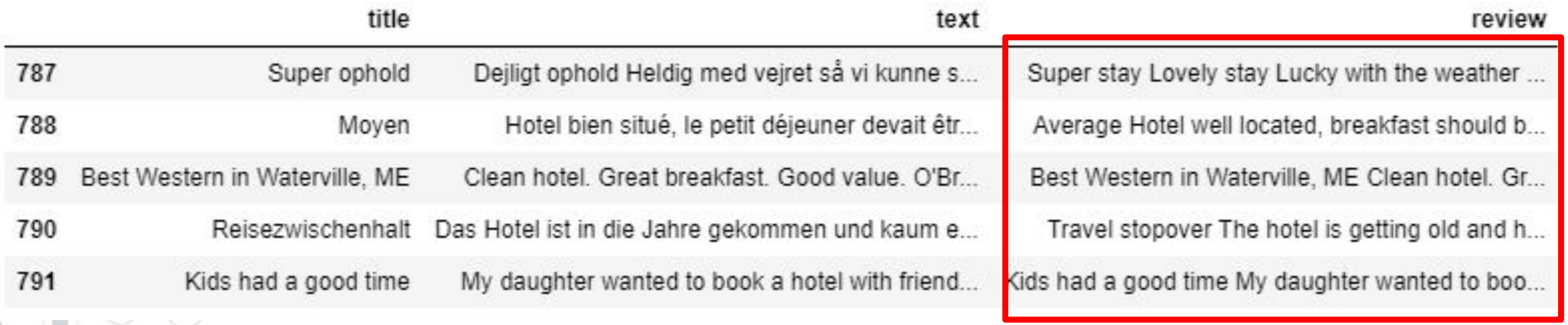


	title	text
184		Came to Binghamton to visit with a relative. F...
185		Great Experience. Nice Staff clean rooms good ...
885		The motel was clean and comfortable bed and a ...
1261		VERY FRIENDLY AND GREAT PRICE. I HAVE STAYED A...
1262		Editorial Review by Citysearch Editors Colonia...
...	...	...
35360		Had the room across from the laundry room. Hea...
35361		clean and comfortable
35592		It was pleasant and near to my family who we w...
40680		Staffs and service were excellent. I will defi...
46512		Bad: Fit and finish construction of rooms coul...

Data Cleaning - Foreign Language Reviews



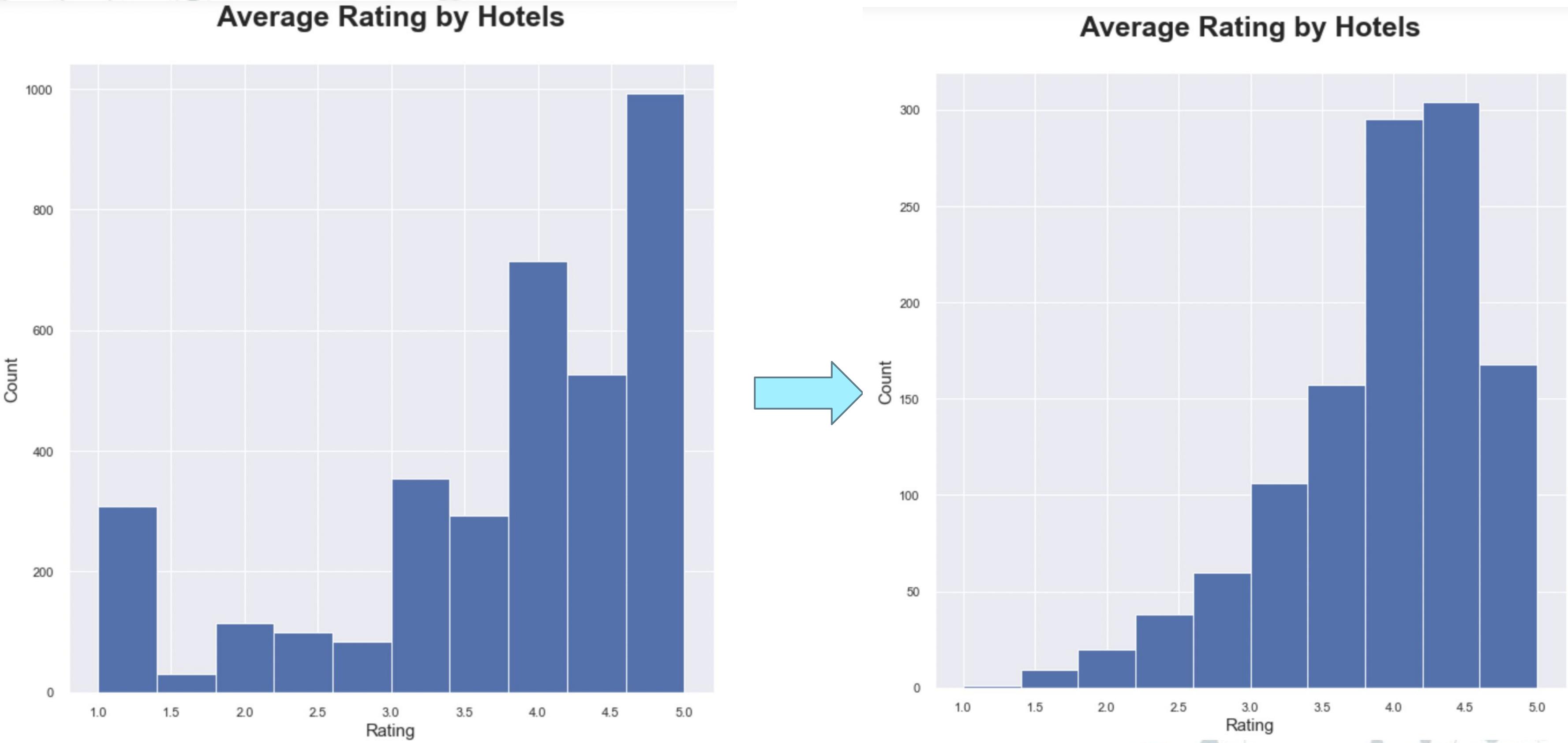
	title	text	review
787	Super ophold	Dejligt ophold Heldig med vejret så vi kunne s...	Super ophold Dejligt ophold Heldig med vejret ...
788	Moyen	Hotel bien situé, le petit déjeuner devait êtr...	Moyen Hotel bien situé, le petit déjeuner deva...
789	Best Western in Waterville, ME	Clean hotel. Great breakfast. Good value. O'Br...	Best Western in Waterville, ME Clean hotel. Gr...
790	Reisezwischenhalt	Das Hotel ist in die Jahre gekommen und kaum e...	Reisezwischenhalt Das Hotel ist in die Jahre g...
791	Kids had a good time	My daughter wanted to book a hotel with friend...	Kids had a good time My daughter wanted to boo...



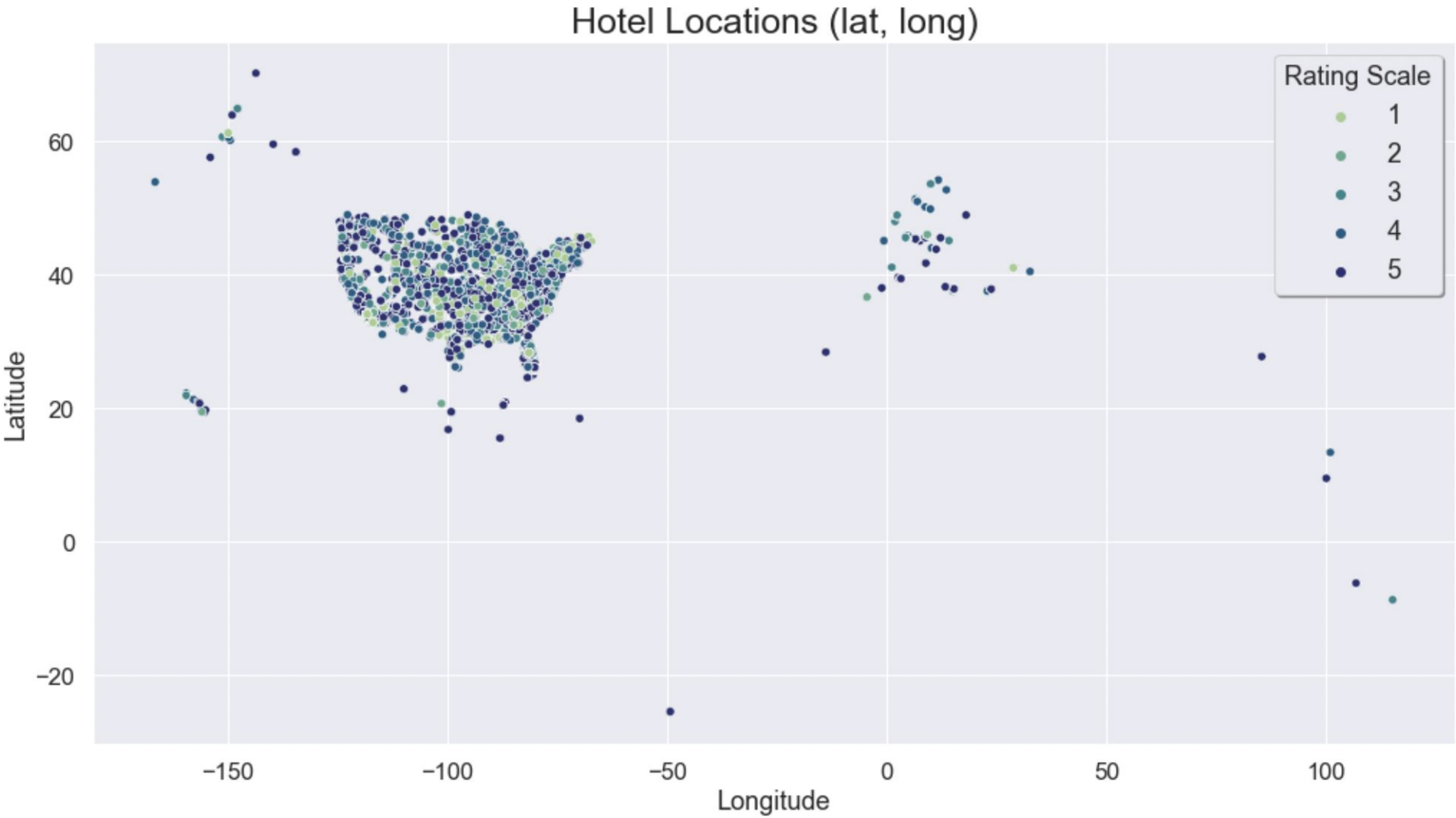
	title	text	review
787	Super ophold	Dejligt ophold Heldig med vejret så vi kunne s...	Super stay Lovely stay Lucky with the weather ...
788	Moyen	Hotel bien situé, le petit déjeuner devait êtr...	Average Hotel well located, breakfast should b...
789	Best Western in Waterville, ME	Clean hotel. Great breakfast. Good value. O'Br...	Best Western in Waterville, ME Clean hotel. Gr...
790	Reisezwischenhalt	Das Hotel ist in die Jahre gekommen und kaum e...	Travel stopover The hotel is getting old and h...
791	Kids had a good time	My daughter wanted to book a hotel with friend...	Kids had a good time My daughter wanted to boo...

3. Exploratory Data Analysis

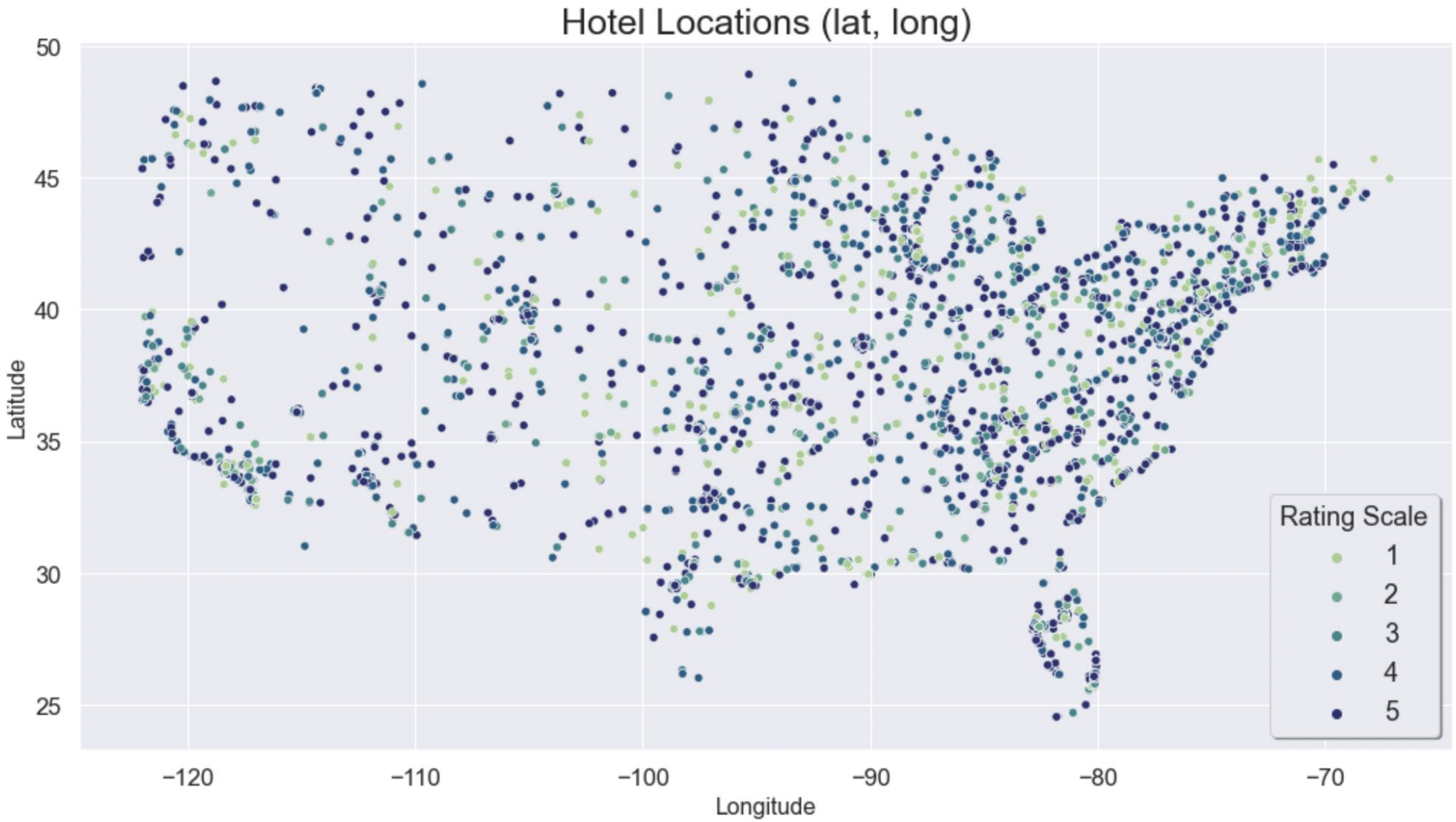
Average Ratings by Hotels



Ratings by Hotel Locations



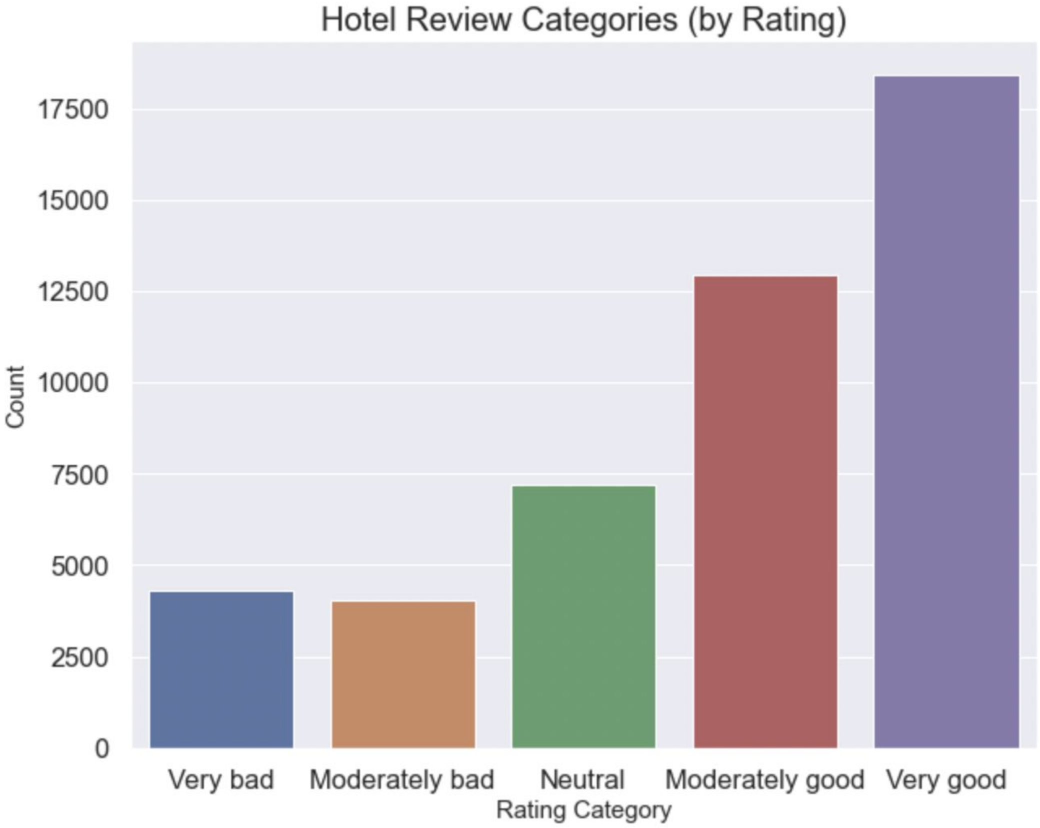
Ratings by Hotel Locations



Ratings by Category (5-class)

Categories based on rounded rating $(int(r))$

- Very good (5)
- Moderately good (4)
- Neutral (3)
- Moderately bad (2)
- Very bad (1)

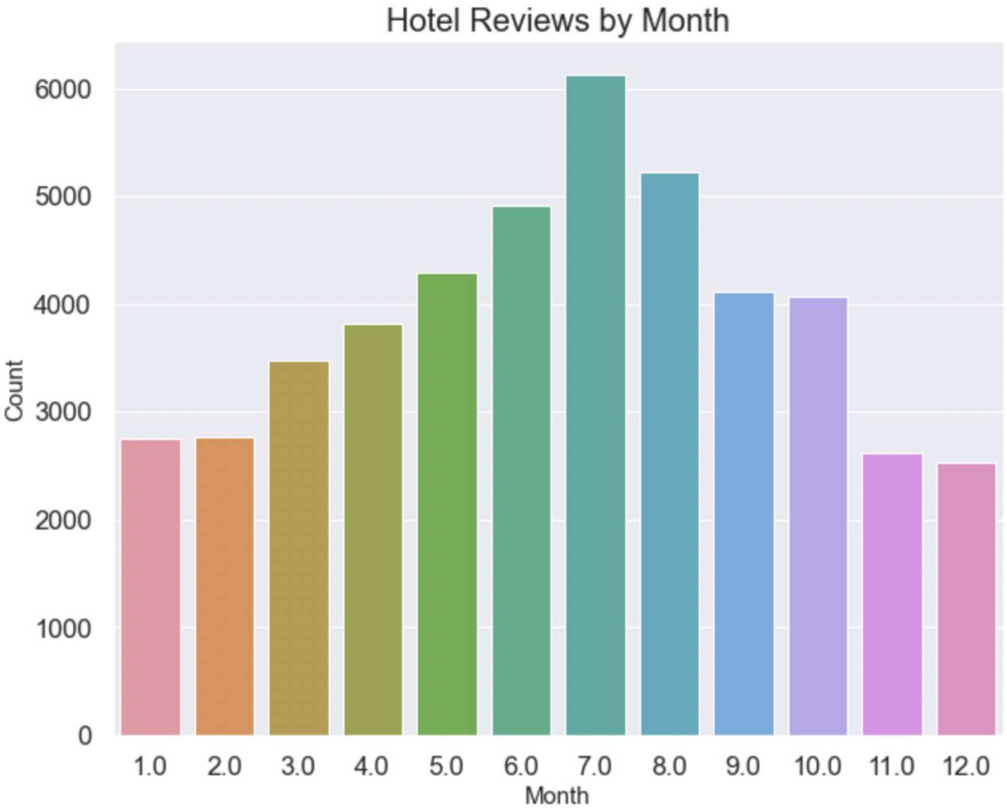


Ratings by Category (3-class)

Categories based on rounded rating $(int(r))$

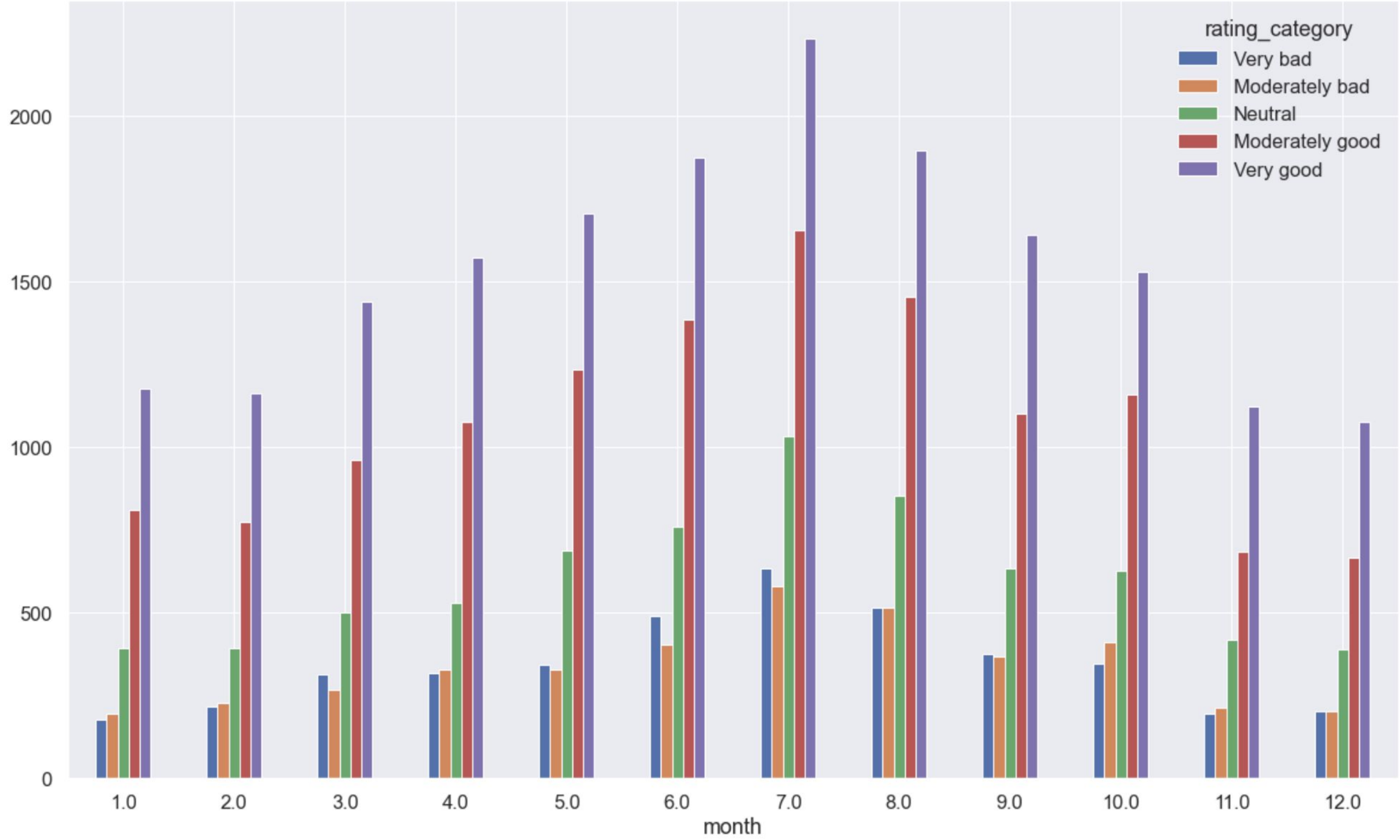
- Very good (5)
- Moderately good (4)
- Neutral (3)
- Moderately bad (2)
- Very bad (1)

Number of Reviews by Month

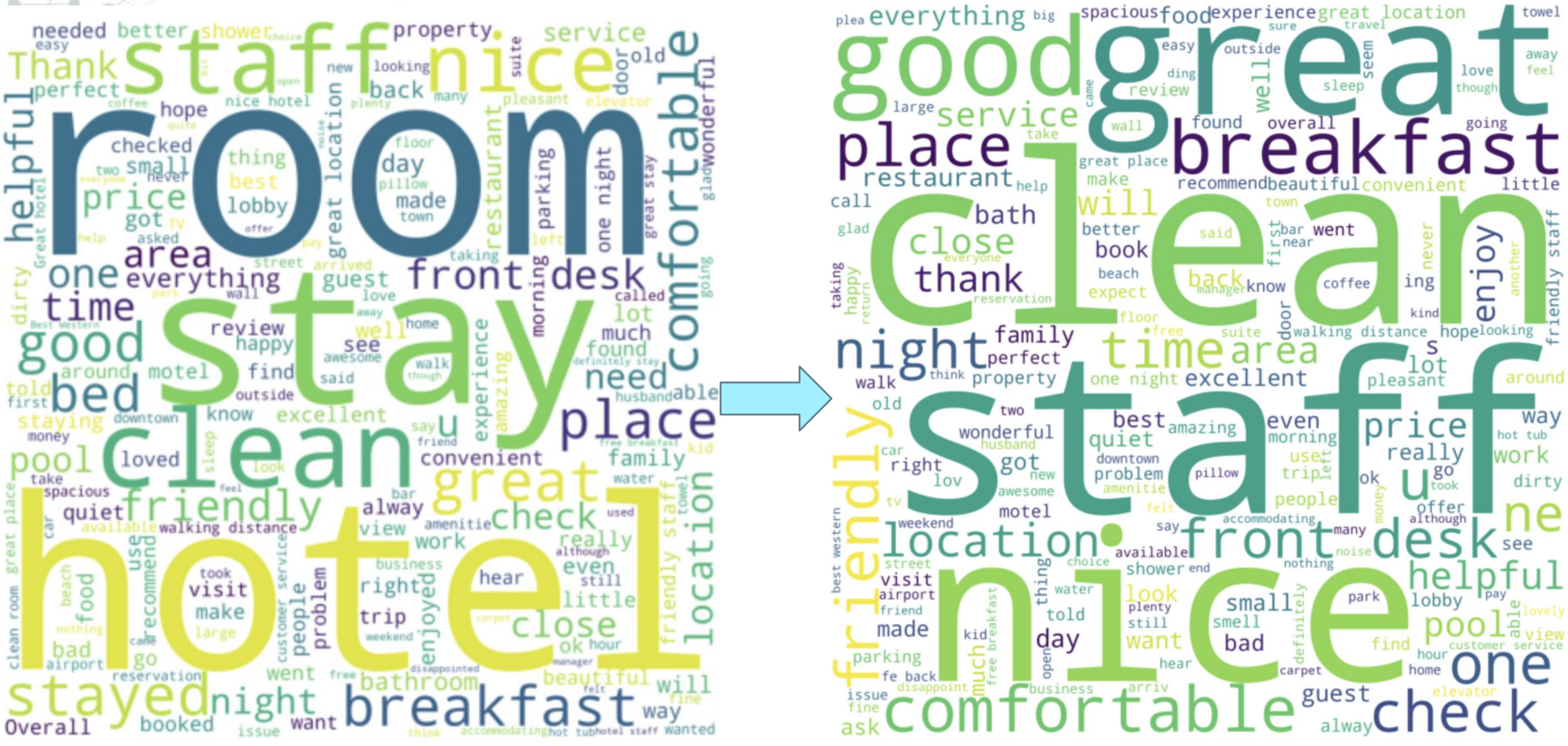


- Summer months have the most reviews

Number of Reviews by Month and Rating Category

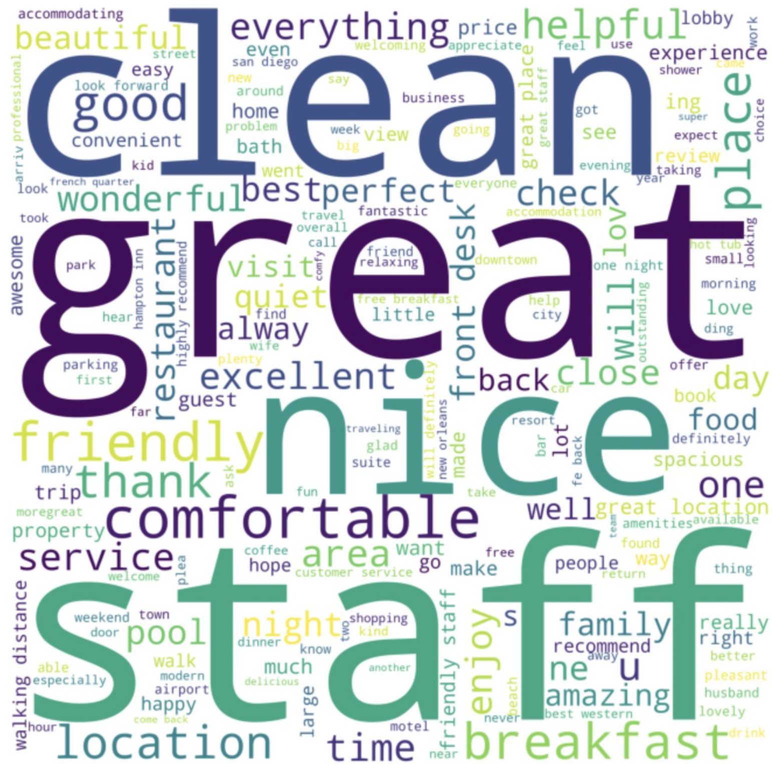


Review WordClouds

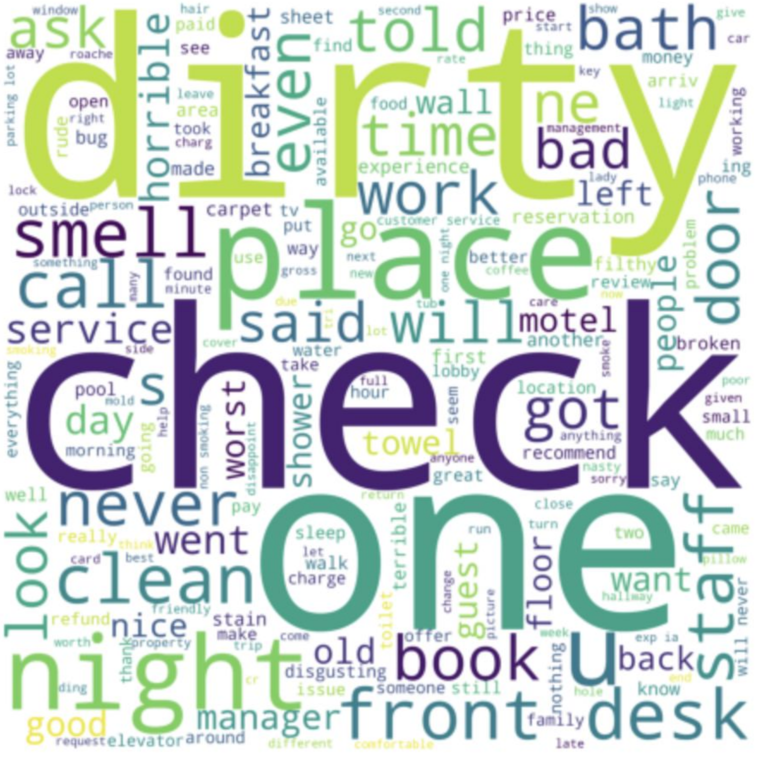


Review WordClouds By Ratings

Very good ratings ($r = 5$)

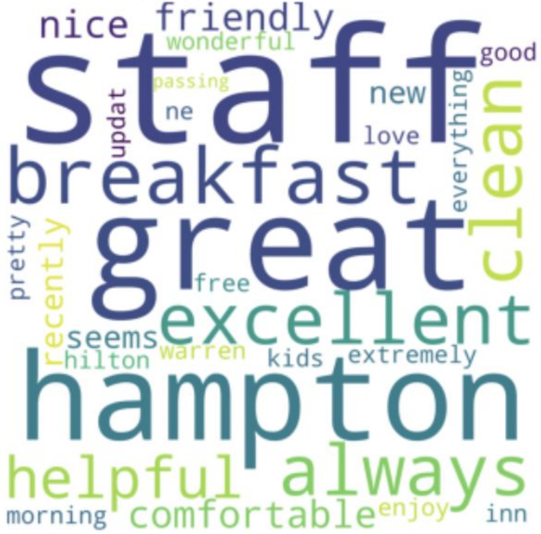


Very bad ratings ($r = 1$)

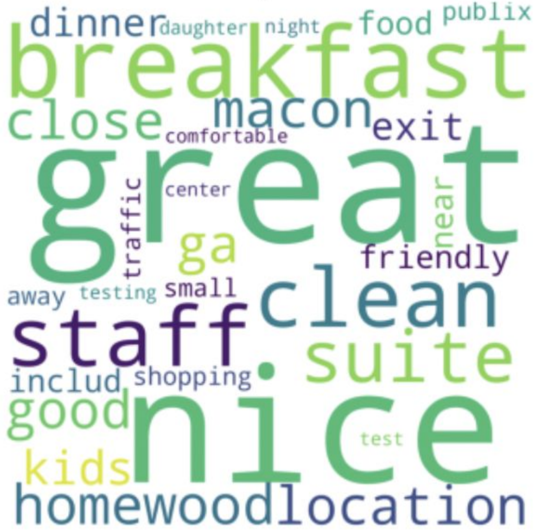


Review WordClouds for Best Hotels

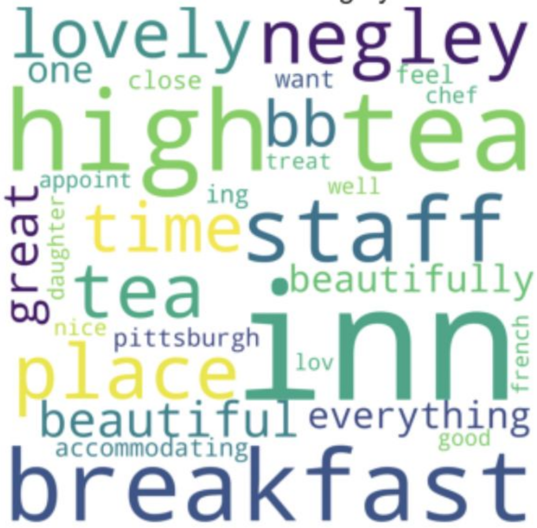
Hampton Inn & Suites Warren



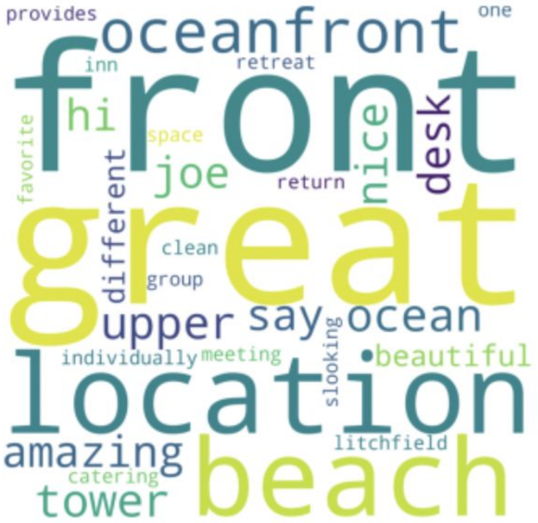
Homewood Suites by Hilton Macon-North

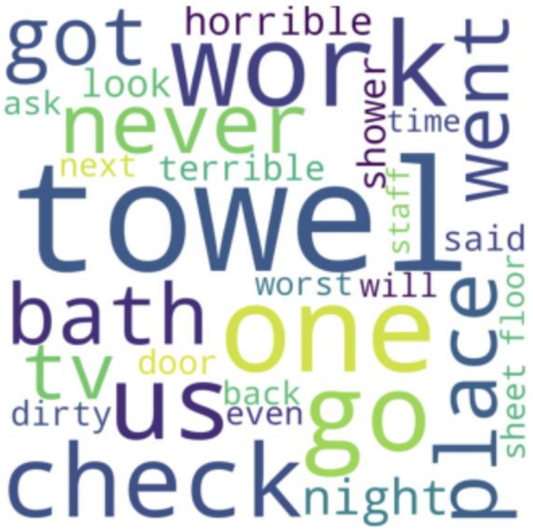


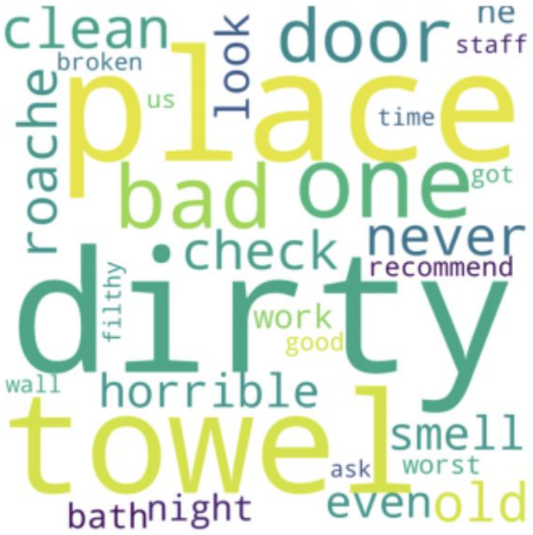
The Inn On Negley



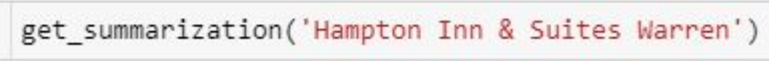
Review WordClouds for Worst Hotels


The Litchfield Inn


Norwood Inn and Suites

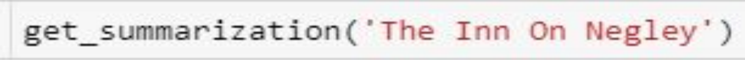

Fiesta Inn and Suites

Review Summarization



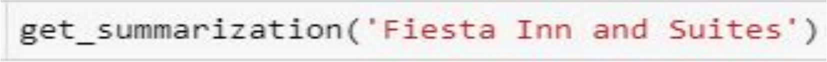
```
get_summarization('Hampton Inn & Suites Warren')
```

the staff is extremely friendly and there's free breakfast in the morning. wonderful beds, very helpful staff and great breakfast.great the staff welcomed me as the guest of the day! the   is nice, seems fairly new, or recently updated, and as we've found at most other hilton brand the staff is wonderful.



```
get_summarization('The Inn On Negley')
```

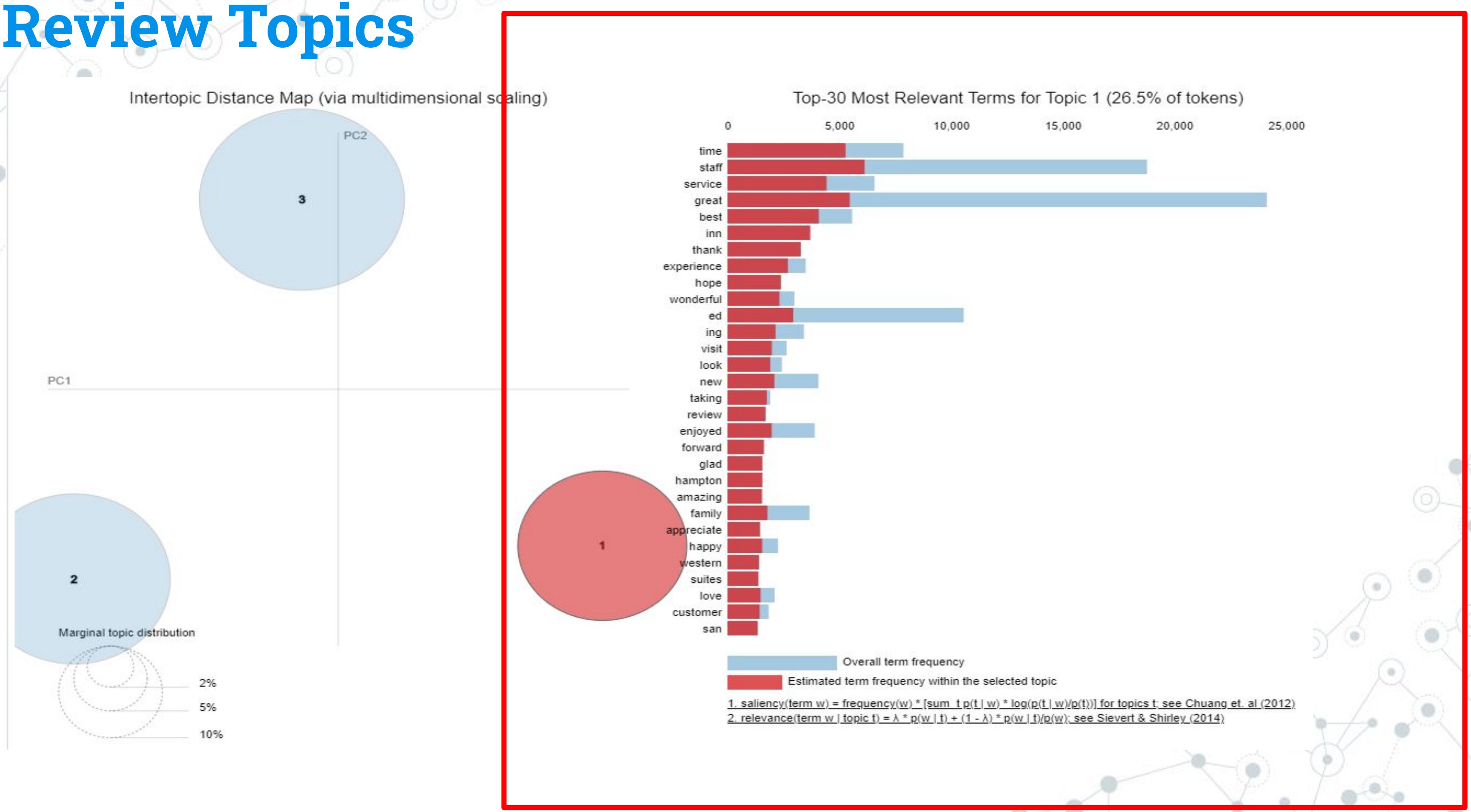
hot breakfast was excellent and hosts very friendly.great inn close to downtown french toast is amazing.a wonderful and relaxing only ed one night, but checked in at 1:00 pm and left at 11:00 the next day so it wasn't just in to sleep and out again.



```
get_summarization('Fiesta Inn and Suites')
```

we struggled to get them to even give us clean towels from the   that they never cleaned, as they said on the 3rd floor they clean it only time week, bad service, at first they told us that we only had reservation for one   and it was not beds as we needed it, the smelled exaggeratedly of tobacco, only the location of the   was good, but overall disaster, will not come back !!!

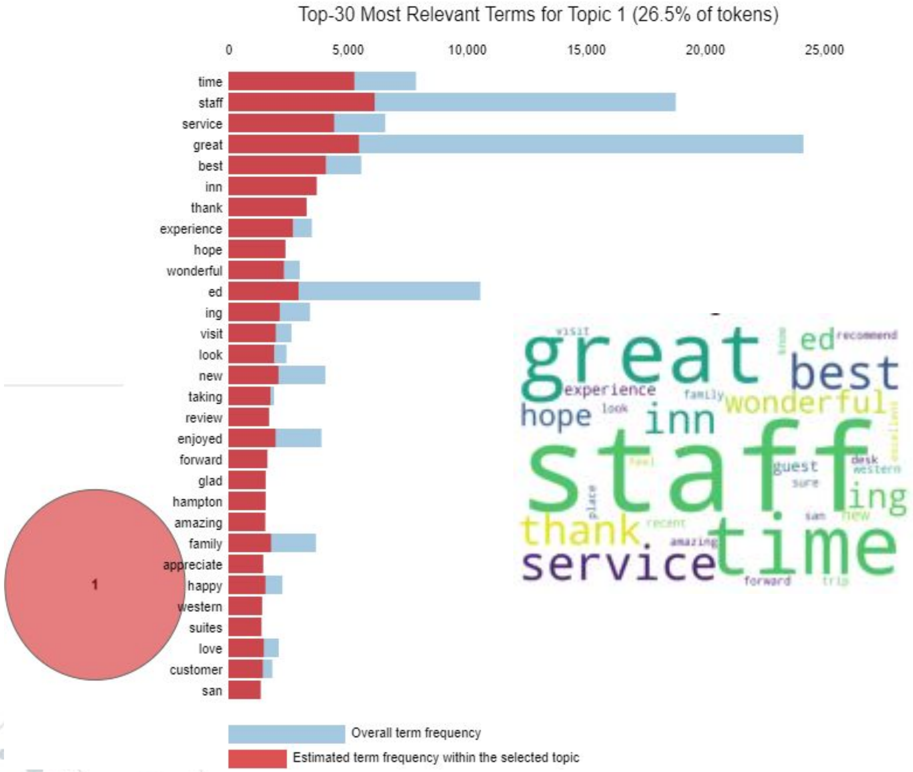
Review Topics



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Review Topics



4. Pipeline Definitions & Evaluation

Feature Engineering

From the texts, we were able to generate two kind of features:

Meta Features (form of text)	Text Features (text content)
● Sentence length (characters & words) ● Word length ● Percentage of unique words ● Stopword count ● Adjective to noun ratio	● Translation of foreign languages ● NRC data analysis: positive/negative ● TF-IDF (words n-grams): the degree to which a sentiment state is related to a word more than the other 2 states

Feature Selection

Reduce number of features

- 50
- 100
- 150
- Not reduced

X

Feature selectors

- Univariate feature selection
- Recursive feature elimination

X

Predictive Models

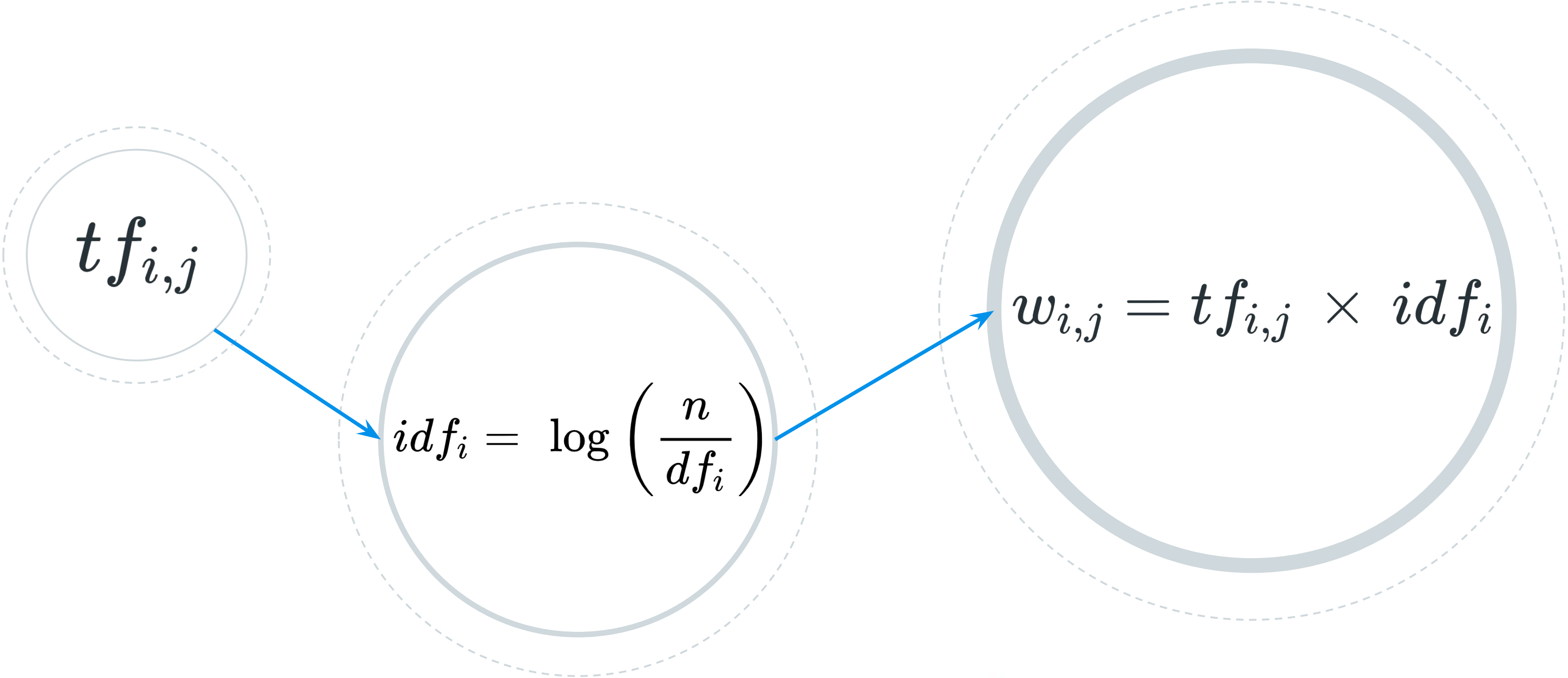
- Decision tree classifier
- Random forest classifier
- Bagging Classifier
- Logistic Regression

4. Predictive Modeling

Modeling Technique Exploration

- ➢ Regression to predict continuous ratings
- ➢ Classification
  - ○ Multiclass Classification on categorized ratings
    - ■ 5-class
    - ■ 3-class
  - ○ Binary Classification

TF-IDF: Term Frequency-Inverse Document Frequency



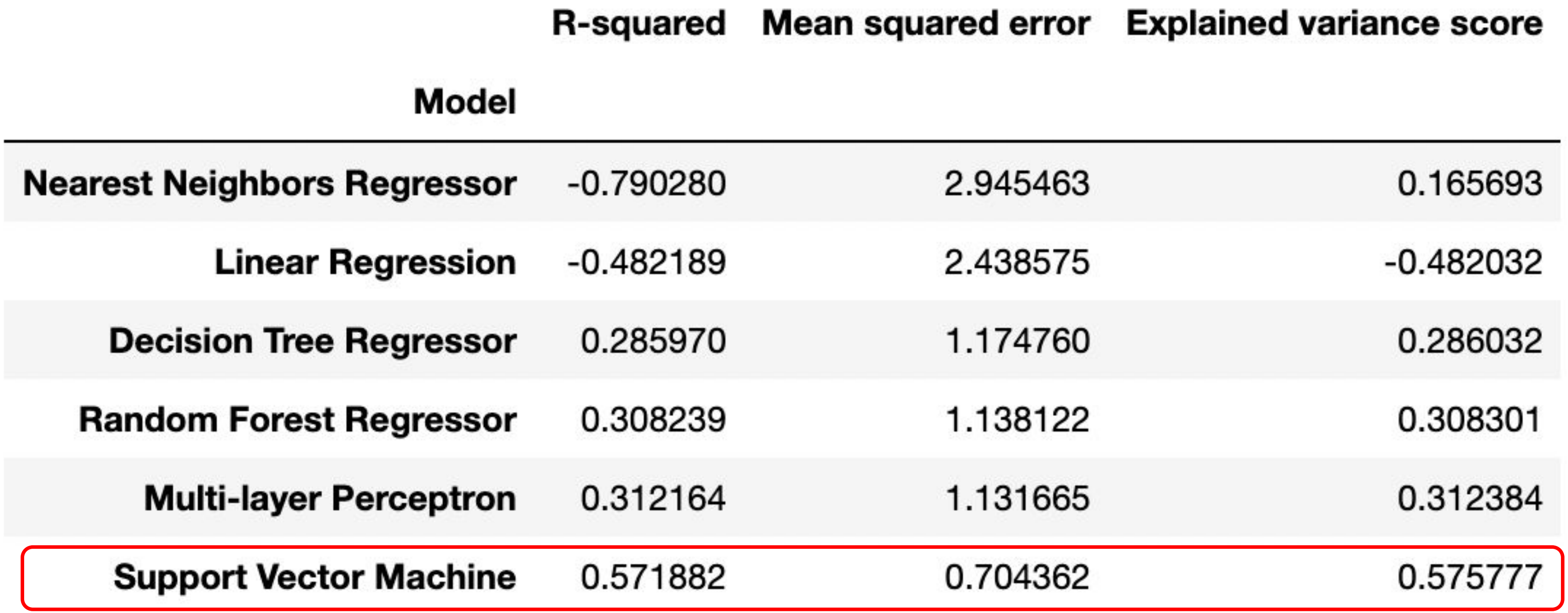
$$tf_{i,j}$$

$$idf_i = \log\left(\frac{n}{df_i}\right)$$

$$w_{i,j} = tf_{i,j} \times idf_i$$

https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530

Regression Models Performance

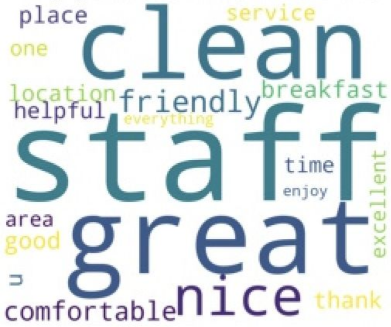


	R-squared	Mean squared error	Explained variance score
Model			
Nearest Neighbors Regressor	-0.790280	2.945463	0.165693
Linear Regression	-0.482189	2.438575	-0.482032
Decision Tree Regressor	0.285970	1.174760	0.286032
Random Forest Regressor	0.308239	1.138122	0.308301
Multi-layer Perceptron	0.312164	1.131665	0.312384
Support Vector Machine	0.571882	0.704362	0.575777

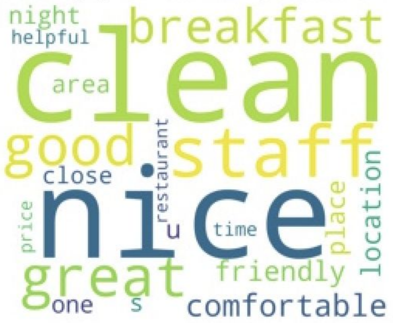
Classification Model Performance

- 5-class classification
- 3-class classification
- Binary classification (2-class)

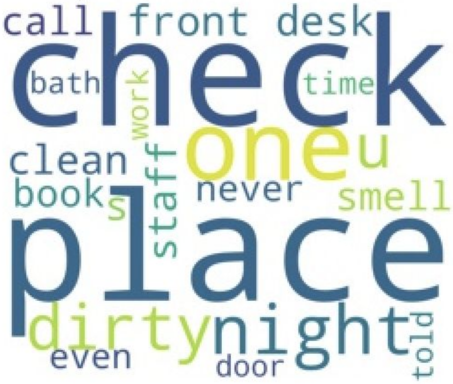
Very-good Reviews $(4.5 \le R \le 5)$



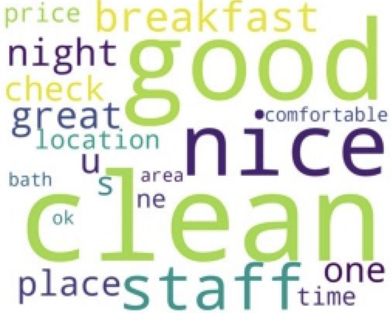
Moderately-good Reviews $(3.5 \le R < 4.5)$



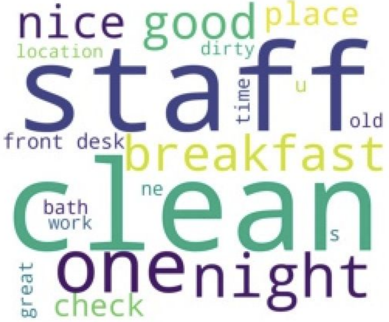
Very-bad Reviews $1 \le R < 1.5$



Neutral Reviews $(2.5 \le R < 3.5)$

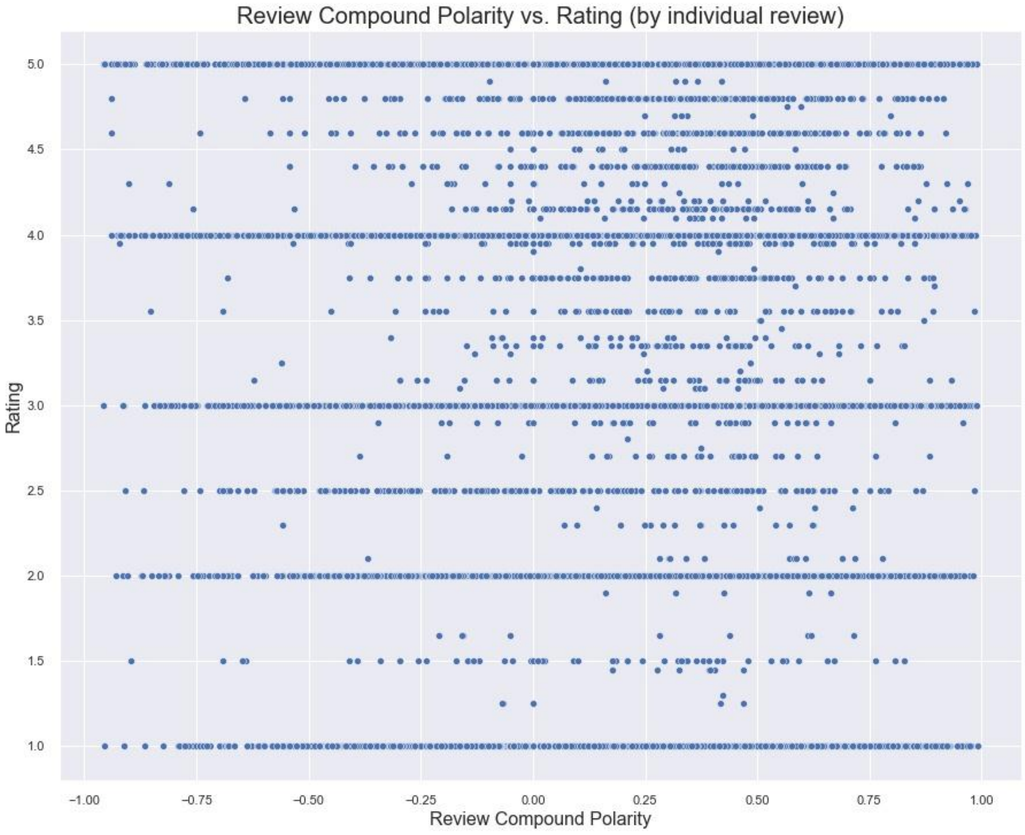


Moderately-bad Reviews $(1.5 \le R < 2.5)$



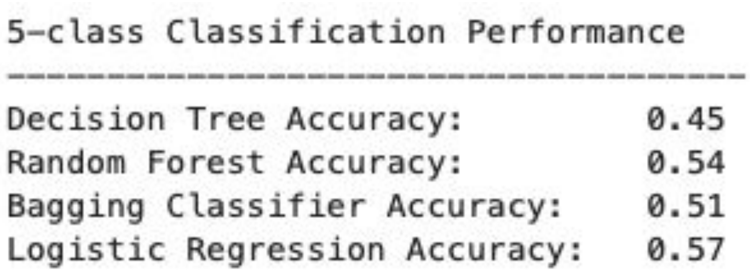
Vader Analysis for Review Text

- Compound sentiment disparity between reviews and their rating



- Low correlation between reviews and their corresponding rating.
- In general, higher compound polarity scores for reviews do not correspond to high ratings, and vice versa.
- There is no clear relationship between review polarity and the corresponding rating.
- 5-class classification models may not exhibit strong performance as a result.

5-class Classification In-sample Model Performance/Comparison*



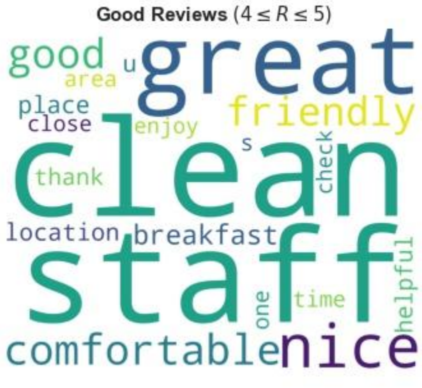
```
5-class Classification Performance
------------------------------------
Decision Tree Accuracy:          0.45
Random Forest Accuracy:          0.54
Bagging Classifier Accuracy:     0.51
Logistic Regression Accuracy:    0.57
```

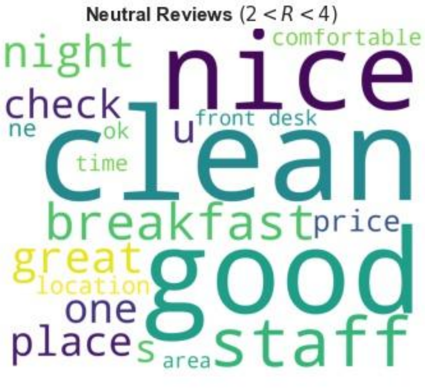
- As expected, model performance for 5-class classification is not very strong.

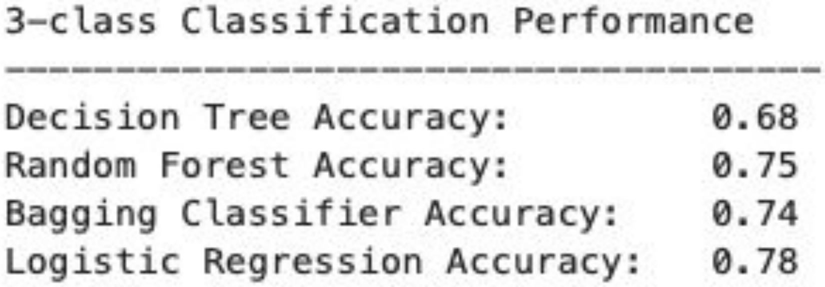
- While our results indicate that logistic regression performs best, an accuracy score of 0.57 is indicative of relatively weak performance (i.e., ability to discriminate between the 5 ratings based on review text is not very strong).

** Performance measured on in-sample test set (80:20 train/test split)*

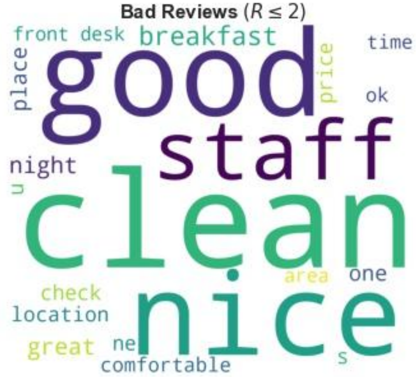
3-class Classification In-Sample Model Performance


Good Reviews $(4 \leq R \leq 5)$

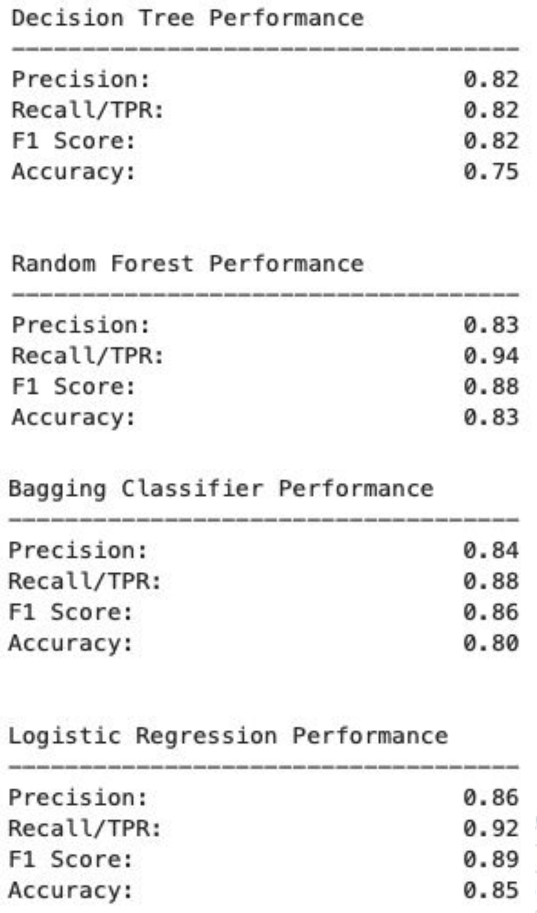

Neutral Reviews $(2 < R < 4)$



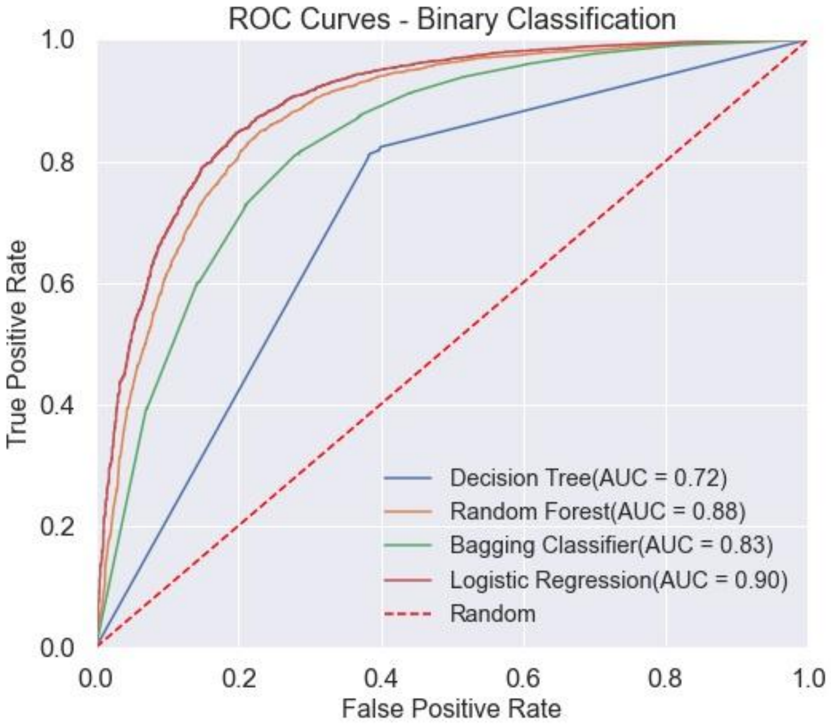
```
3-class Classification Performance
----------------------------------------
Decision Tree Accuracy:           0.68
Random Forest Accuracy:           0.75
Bagging Classifier Accuracy:      0.74
Logistic Regression Accuracy:     0.78
```


Bad Reviews $(R \leq 2)$

Binary Classifier In-sample Model Performance/Comparison* (review text only)

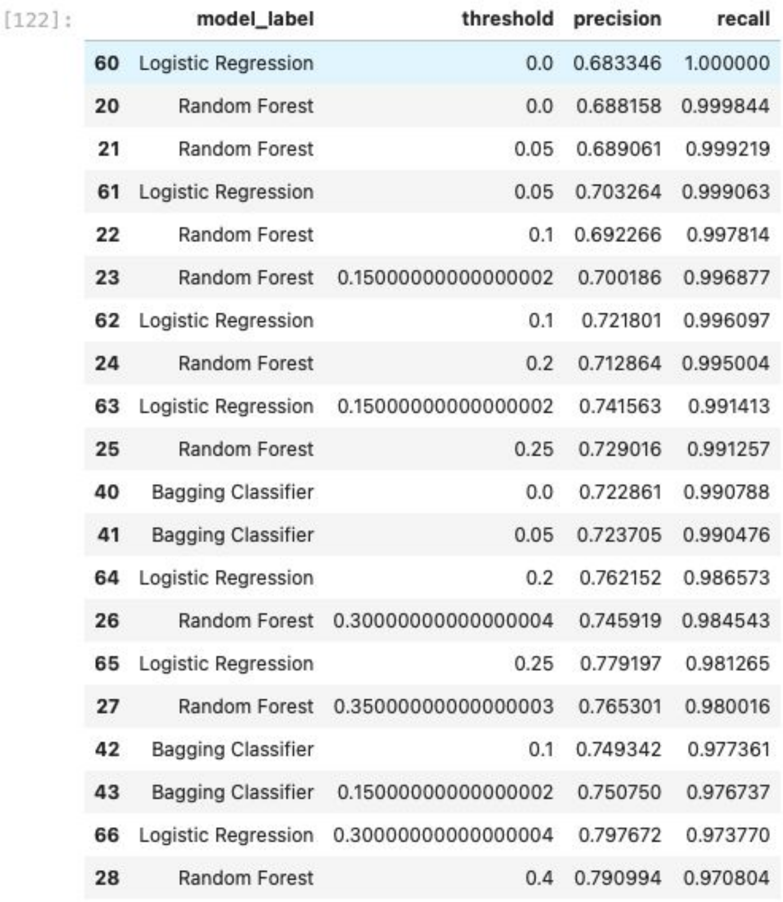


```
Decision Tree Performance
------------------------------------
Precision:                      0.82
Recall/TPR:                     0.82
F1 Score:                       0.82
Accuracy:                       0.75


Random Forest Performance
------------------------------------
Precision:                      0.83
Recall/TPR:                     0.94
F1 Score:                       0.88
Accuracy:                       0.83

Bagging Classifier Performance
------------------------------------
Precision:                      0.84
Recall/TPR:                     0.88
F1 Score:                       0.86
Accuracy:                       0.80


Logistic Regression Performance
------------------------------------
Precision:                      0.86
Recall/TPR:                     0.92
F1 Score:                       0.89
Accuracy:                       0.85
```



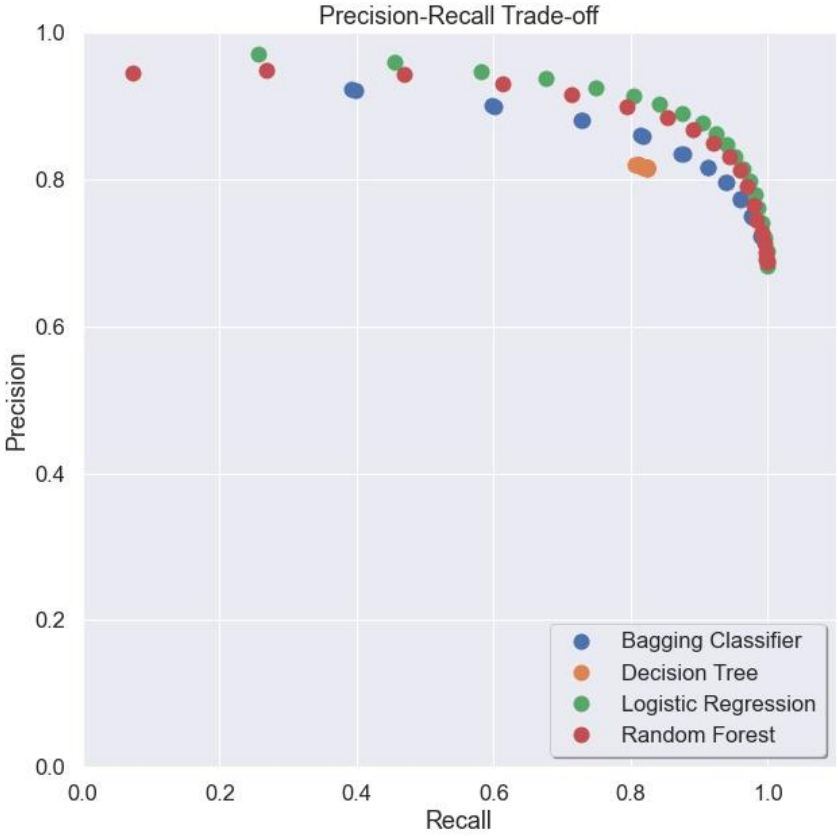
** Performance measured on in-sample test set (80:20 train/test split)*

Precision-Recall Tradeoff for Binary Classifiers



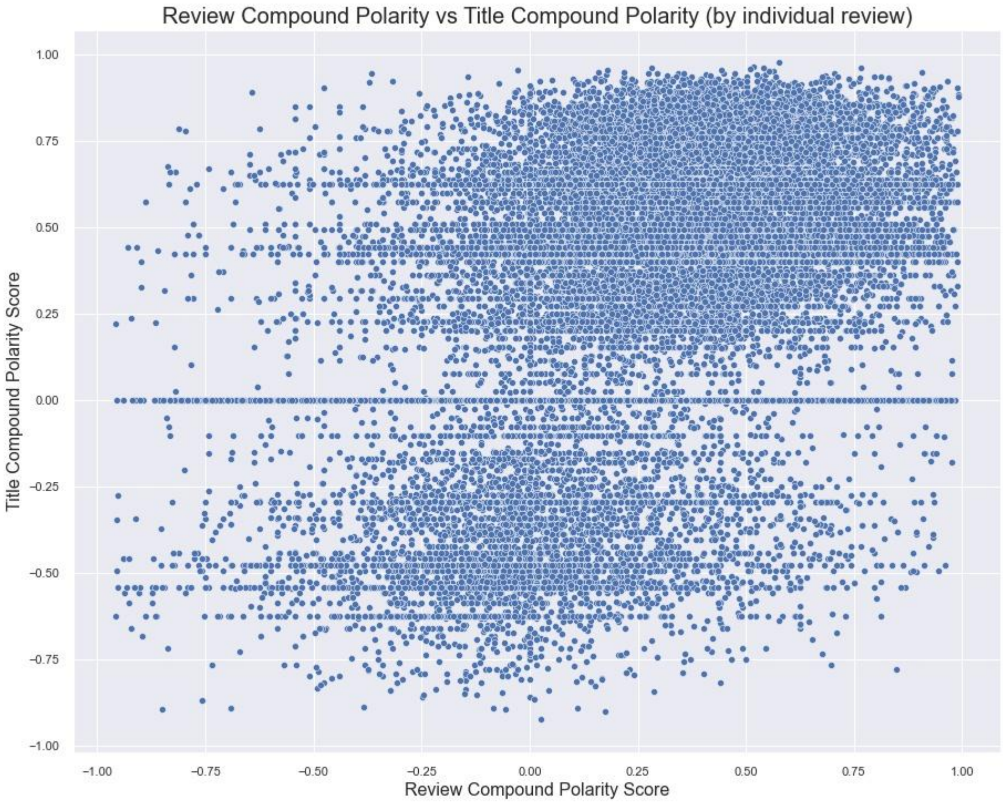
[122]:

	model_label	threshold	precision	recall
60	Logistic Regression	0.0	0.683346	1.000000
20	Random Forest	0.0	0.688158	0.999844
21	Random Forest	0.05	0.689061	0.999219
61	Logistic Regression	0.05	0.703264	0.999063
22	Random Forest	0.1	0.692266	0.997814
23	Random Forest	0.15000000000000002	0.700186	0.996877
62	Logistic Regression	0.1	0.721801	0.996097
24	Random Forest	0.2	0.712864	0.995004
63	Logistic Regression	0.15000000000000002	0.741563	0.991413
25	Random Forest	0.25	0.729016	0.991257
40	Bagging Classifier	0.0	0.722861	0.990788
41	Bagging Classifier	0.05	0.723705	0.990476
64	Logistic Regression	0.2	0.762152	0.986573
26	Random Forest	0.30000000000000004	0.745919	0.984543
65	Logistic Regression	0.25	0.779197	0.981265
27	Random Forest	0.35000000000000003	0.765301	0.980016
42	Bagging Classifier	0.1	0.749342	0.977361
43	Bagging Classifier	0.15000000000000002	0.750750	0.976737
66	Logistic Regression	0.30000000000000004	0.797672	0.973770
28	Random Forest	0.4	0.790994	0.970804



More Vader Analysis

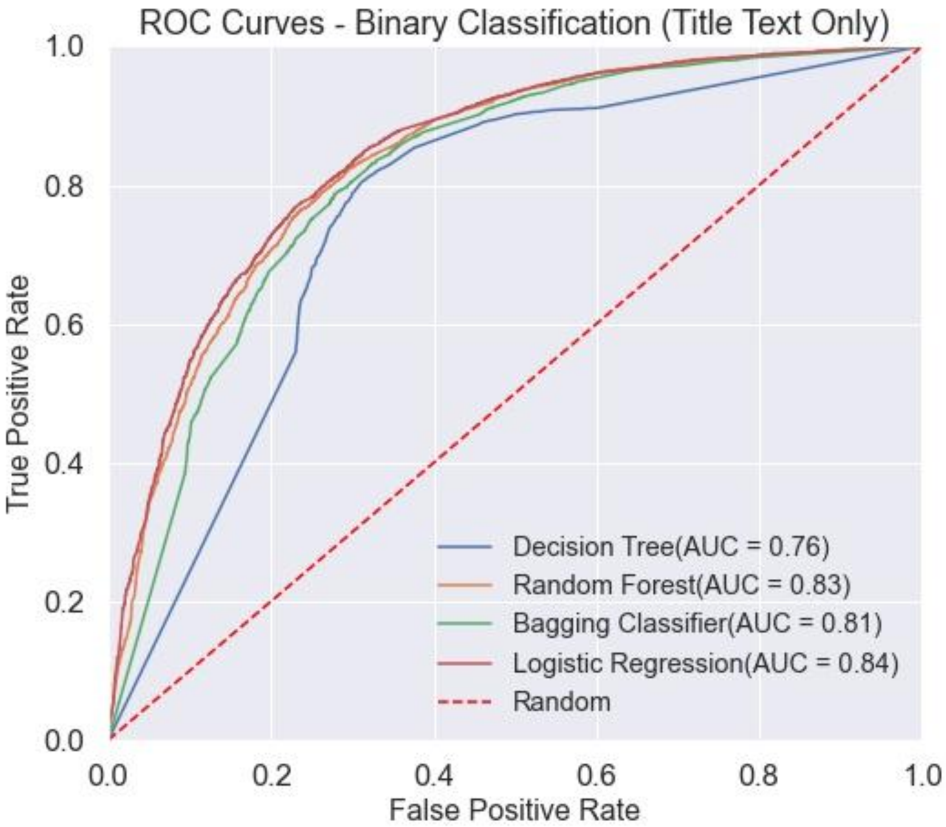
- Compound polarity of review text vs. compound polarity of title text



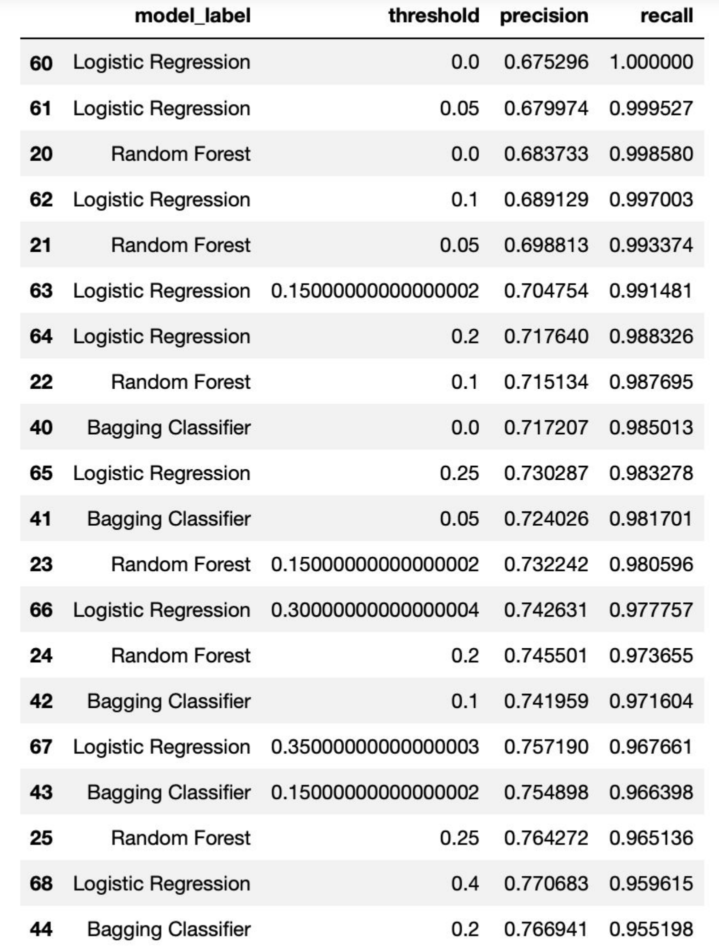
- Compound sentiment disparity between review text and their correspond title text.
- Low correlation between compound scores for review text and title text.
- Should we model them separately? Together? Only consider reviews?

Predicting ratings using title text only (excluding review text)

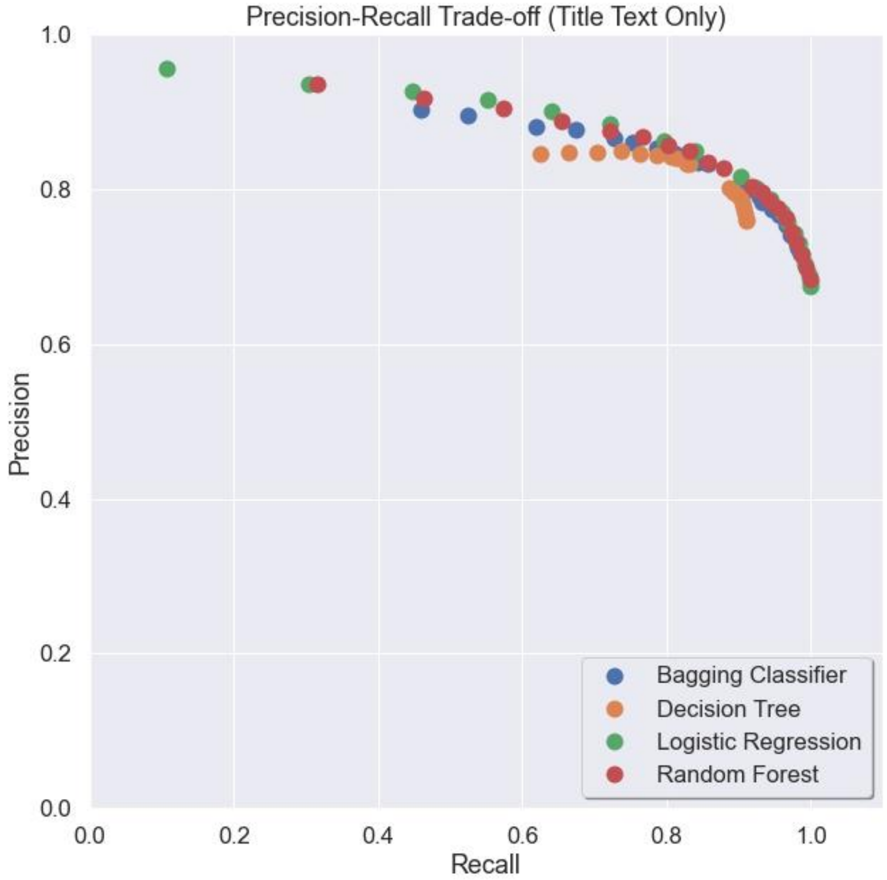
```
Decision Tree Performance
------------------------------------
Precision:                      0.83
Recall/TPR:                     0.83
F1 Score:                       0.83
Accuracy:                       0.77


Random Forest Performance
------------------------------------
Precision:                      0.83
Recall/TPR:                     0.88
F1 Score:                       0.85
Accuracy:                       0.80


Bagging Classifier Performance
------------------------------------
Precision:                      0.83
Recall/TPR:                     0.86
F1 Score:                       0.84
Accuracy:                       0.79


Logistic Regression Performance
------------------------------------
Precision:                      0.80
Recall/TPR:                     0.93
F1 Score:                       0.86
Accuracy:                       0.80
```



Precision-Recall on title text only



	model_label	threshold	precision	recall
60	Logistic Regression	0.0	0.675296	1.000000
61	Logistic Regression	0.05	0.679974	0.999527
20	Random Forest	0.0	0.683733	0.998580
62	Logistic Regression	0.1	0.689129	0.997003
21	Random Forest	0.05	0.698813	0.993374
63	Logistic Regression	0.15000000000000002	0.704754	0.991481
64	Logistic Regression	0.2	0.717640	0.988326
22	Random Forest	0.1	0.715134	0.987695
40	Bagging Classifier	0.0	0.717207	0.985013
65	Logistic Regression	0.25	0.730287	0.983278
41	Bagging Classifier	0.05	0.724026	0.981701
23	Random Forest	0.15000000000000002	0.732242	0.980596
66	Logistic Regression	0.30000000000000004	0.742631	0.977757
24	Random Forest	0.2	0.745501	0.973655
42	Bagging Classifier	0.1	0.741959	0.971604
67	Logistic Regression	0.35000000000000003	0.757190	0.967661
43	Bagging Classifier	0.15000000000000002	0.754898	0.966398
25	Random Forest	0.25	0.764272	0.965136
68	Logistic Regression	0.4	0.770683	0.959615
44	Bagging Classifier	0.2	0.766941	0.955198



5. Future Work

- Segmentation:
  - geographic location, and/or
  - season
- Review summary for hotels recommendation

Thanks!

Any questions?