HOTEL CANCELLATION PREDICTION

BUSINESS CHALLENGE

- This dataset contains relevant data from hotels in the travel industry

- The goal is to observe hotel cancellation tendencies in travel business to ascertain patterns and predict reservation importance in customer behavior.

Data Pre-Processing

```
df = pd.read_csv('./distorted_data_extra_field.csv', sep=',')
```

```
df.drop(df.columns[[0]], axis=1, inplace=True)
```

```
roomCount = df.loc[df['roomCount'] <= 0 ]

print(len(roomCount))
```

```
3
```

```
cleanRC = df[df['roomCount'] <= 0].index.tolist()
df.drop(df.index[cleanRC], inplace=True)
```

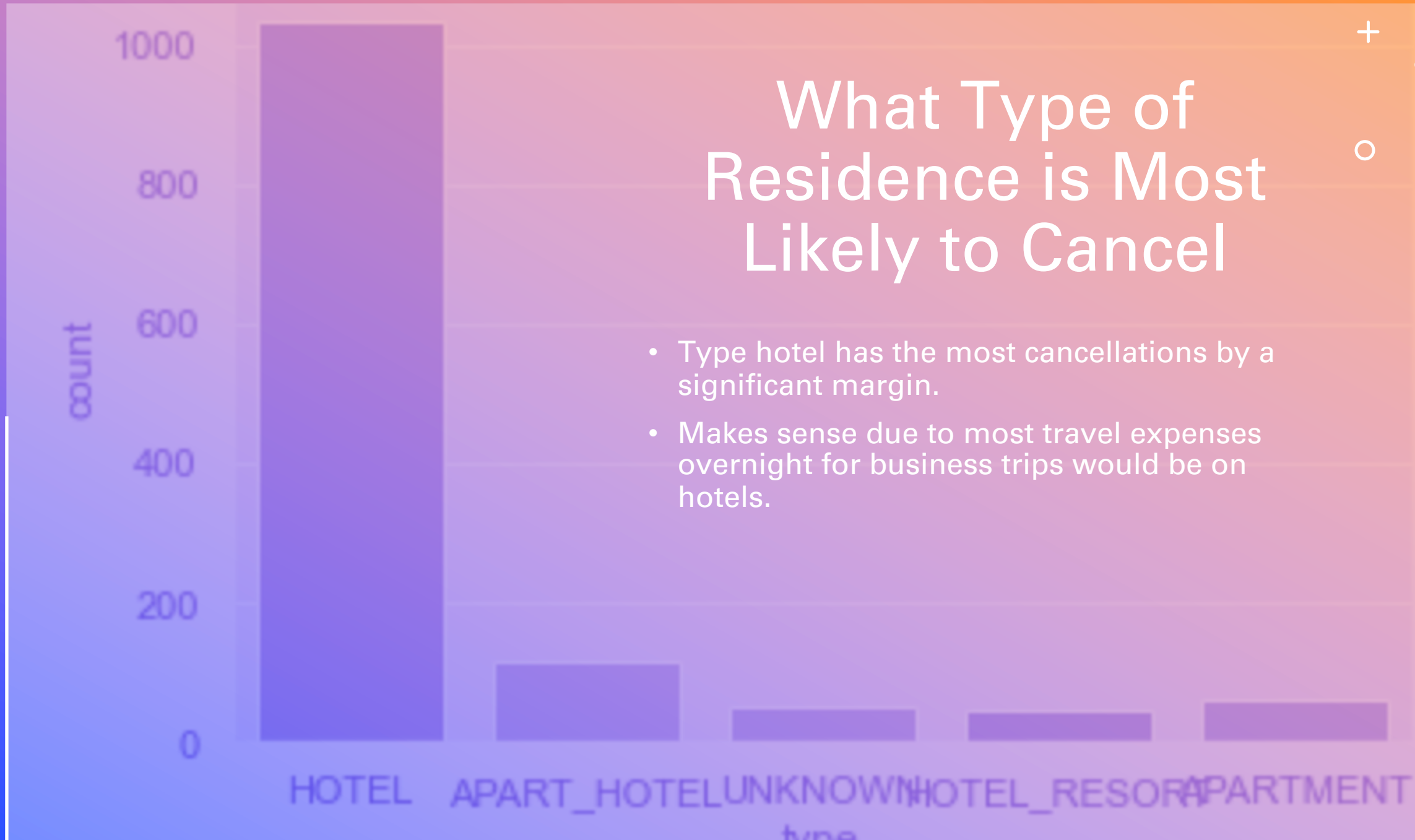
- Remove unnecessary columns
- Handle data anomalies
- There was only one column in the data-set with anomalous data.
- Handled by removing all instances where there were orders with no room counts.

Data Exploration

FIND OUT WHICH VARIABLES MOST INFLUENCE HOTEL CANCELLATION

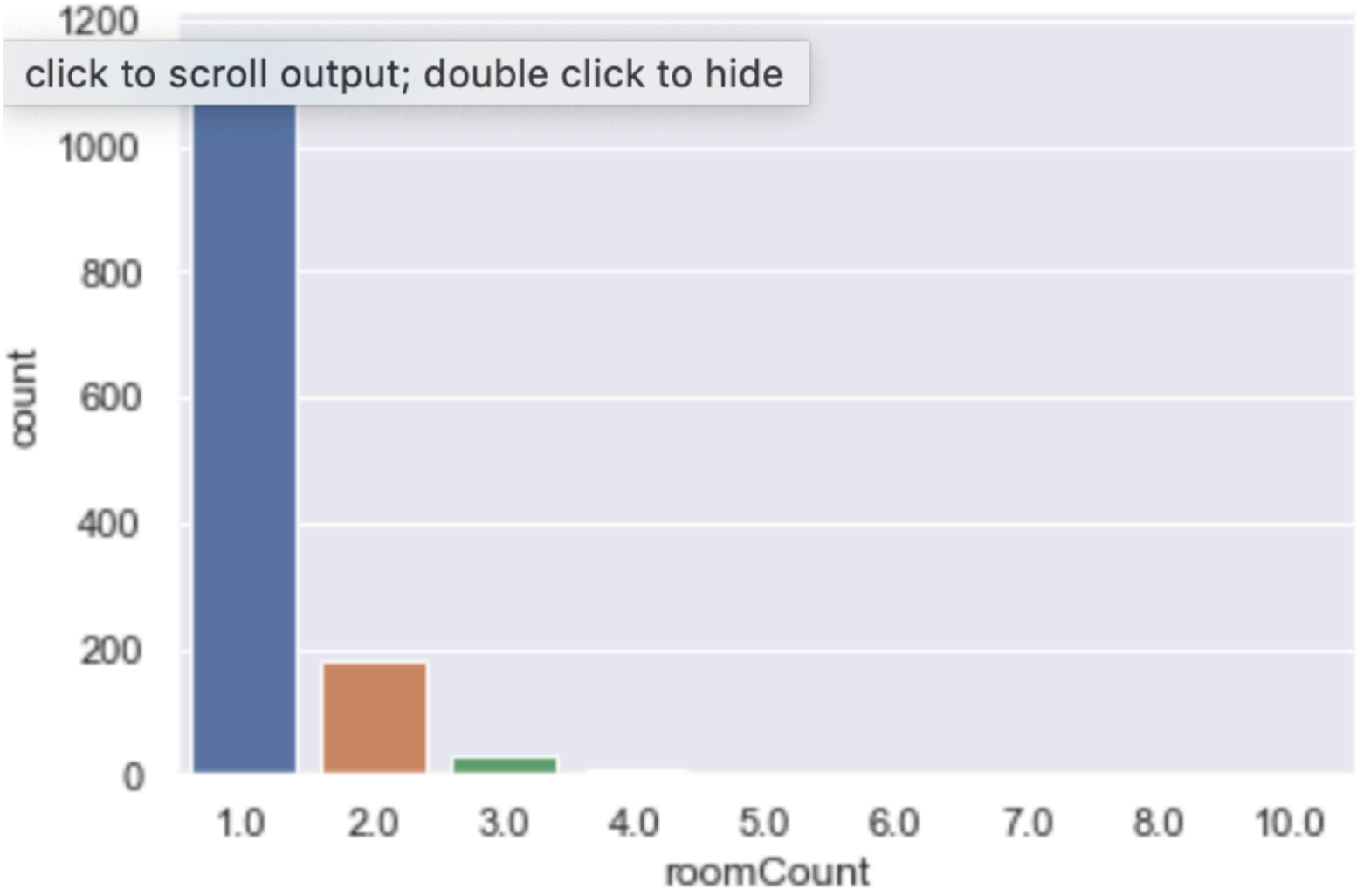
USING VARIOUS FACTORS, VISUALIZE THE FACTORS THAT INFLUENCE HOTEL CANCELLATION

What Type of Residence is Most Likely to Cancel



- Type hotel has the most cancellations by a significant margin.
- Makes sense due to most travel expenses overnight for business trips would be on hotels.

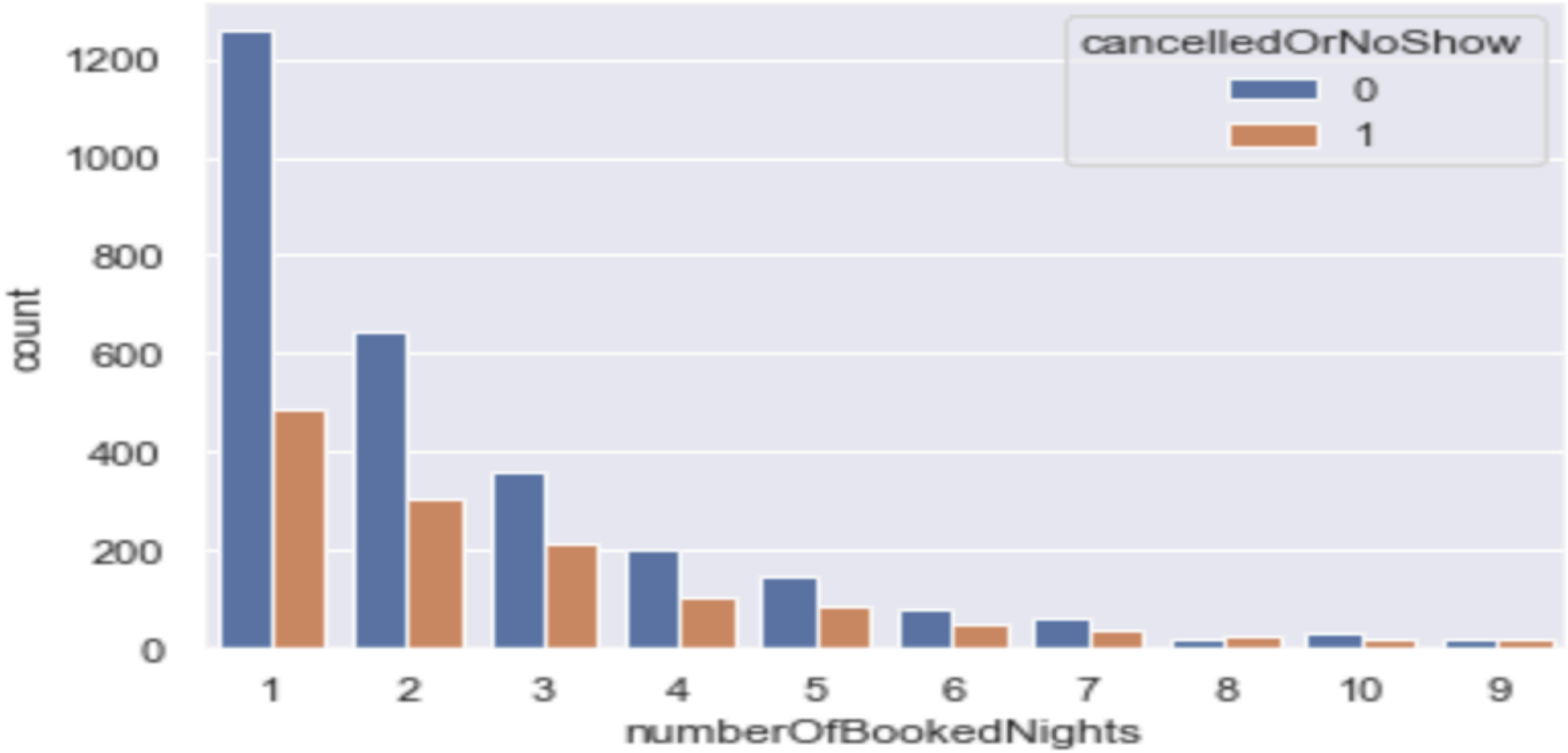
Which Room Count Received The Most Cancelled Orders

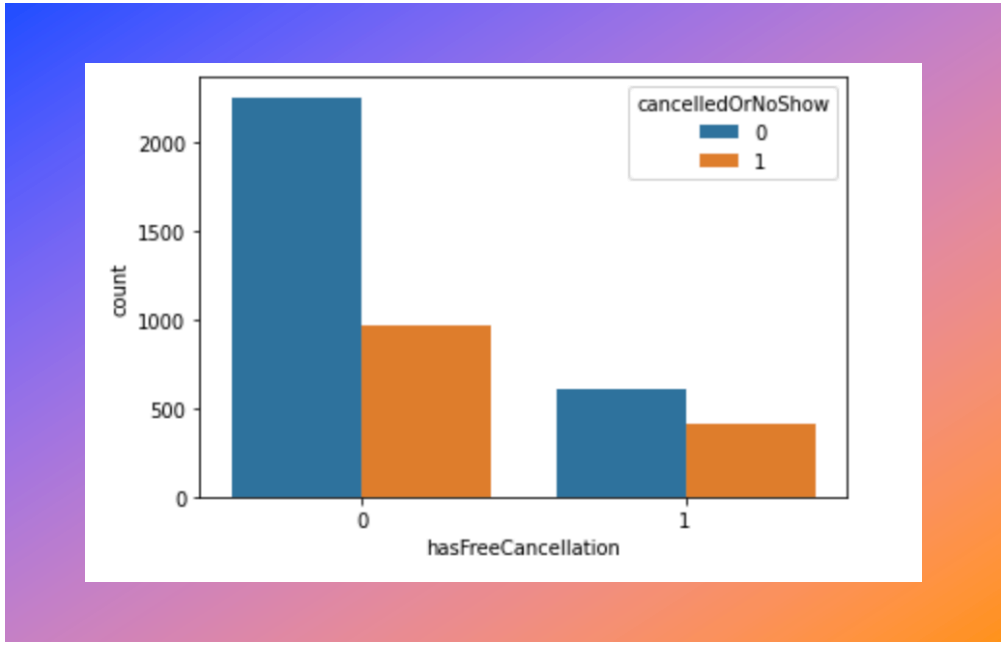


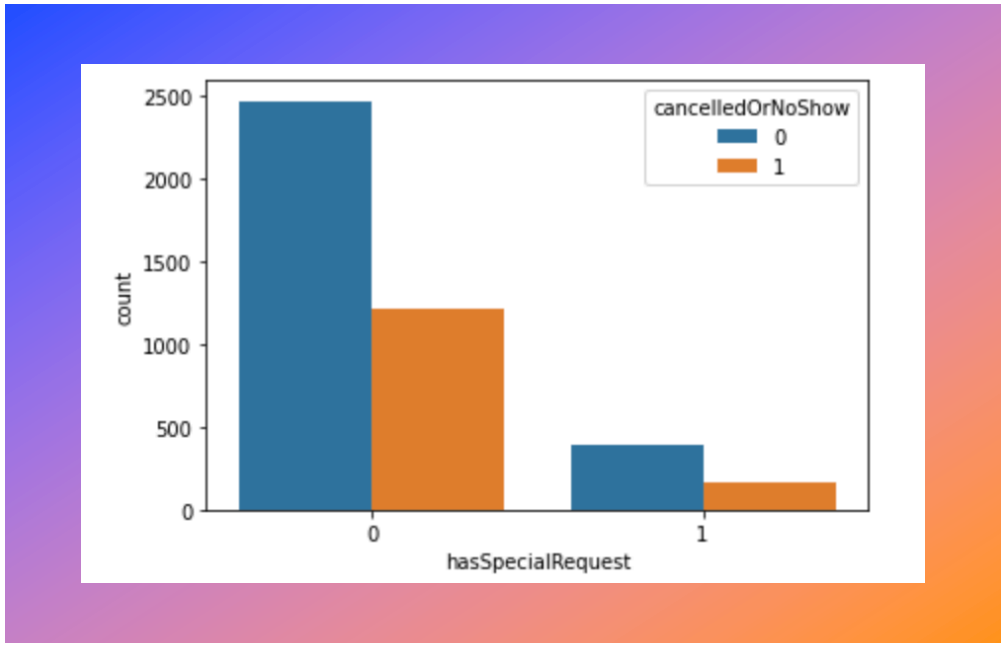
Which Ratings Received The Least Cancelled Orders

- The average user rating of 0 received the most cancelled orders by a significant margin

Is There A Relationship Between Canceled Orders And Number of Booked Nights?







Is There A Relationship Between Special Requests And Canceled Orders?

- Yes, there's a relationship
- Clients that have a free cancellation or a special request are much less likely to cancel their stay

Data Modeling

USE VARIOUS MACHINE LEARNING TECHNIQUES TO HELP FIND WHICH MODEL HELPS BEST PREDICT FUTURE CANCELLATIONS.

BASED ON FINDINGS OF EDA, ALL 3 MODELS WILL ONLY USE SELECT COLUMNS INSTEAD OF ALL THE COLUMNS

THIS IS SO THAT THE MODEL WILL GIVE A MORE ACCURATE INDICATION OF FUTURE CANCELLED RESERVATIONS

Synthetic Minority Oversampling Technique (SMOTE)

```
oversample = SMOTE()
X_train, Y_train = oversample.fit_resample(X_train, Y_train)
counter = Counter(Y_train)
print(counter)
```

Data is imbalanced with a few categories such as 5 star hotels only having a few records

Problem with imbalanced data is that it produces biased predictions

SMOTE will help multiply an instance by a random n amount to help account for the difference between neighboring incidents

Logistic Regression

```
X_train, X_test, Y_train, Y_test = sklearn.cross_validation.train_test_split(df_feat_selected,df['cancelledOrNoShow'],test_size=0.30, random_state=101)
```

```
logreg.fit(X_train, Y_train)

Y_pred = logreg.predict(X_test)

logreg.score(X_test, Y_test)
```

```
0.66719118804091271
```

- An algorithm for classification
- Helps find relationships between features and a particular outcome
- With a score of 0.667, my model is semi-accurate, and it implies that my predictor variables are good indicator of future hotel cancellations

Random Forest Classifier

	precision	recall	f1-score	support
0	0.70	0.84	0.76	863
1	0.41	0.23	0.29	408
avg / total	0.61	0.65	0.61	1271

- Helps select random samples from a dataset
- Using each sample, constructs a decision tree
- Then performs a vote for each sample with an associated predicted result
- Hotels will want to identify customers who are ultimately going to cancel their booking with greater accuracy, thus recall is a better measurement to use
- My recall score is 0.65. Mildly, but not thoroughly accurate

Decision Trees

	precision	recall	f1-score	support
0	0.70	0.87	0.78	863
1	0.45	0.23	0.30	408
avg / total	0.62	0.66	0.62	1271

- Supervised, non-parametric learning method that is used for classification
- Helps create a model that is inferred from the data features by predicting the value of a selected variable through simple decision rules
-  Will use recall over precision as hotels will want want to know which customers will ultimately cancel their orders
-  With a recall score 0.66, my model could be better, but it's usable

Comparison of Models & Conclusion

Used three machine learning models: Logistic Regression, Random Forest, and Decision Trees

None of them were particularly accurate, but do lend credence on the fact that the number of booked nights, the average user rating, and the room count can help predict and identify future cancellation trends

Accuracy for all 3 models in the 60s, but Logistic Regression was the most accurate with 0.67.

A more accurate model could have been accrued if columns were combined.

THANK YOU!