Hotel Reservation Data Analysis
Technical Presentation

Date: April 3, 2020

Presenter: Justin Albright

Can hospitality managers predict guest cancellations in order to optimize revenue?

!

Cancellations have been a longstanding yet unstudied problem in the hospitality industry, which has traditionally been considered beyond the operator's control.

If we study reservation data we can identify trends and patterns in the customer base that cancels its reservations. This information can be used by managers to devise ways to hedge against anticipated future cancellations.

!

With so many unique geographies and hospitality experiences, it can be challenging to identify where to start and what to analyze. Much of this information is proprietary and confidential for business purposes.

Detailed information from nearly 120,000 reservations is available from two hotels in Portugal spanning over two years (July 2015 – August 2017). 32 distinct pieces of data (variables) were collected from each reservation, some numeric and some categorical. Names and other personally identifying information were removed to ensure rights to privacy.

!

How can this be analyzed?

Statistical analysis, both descriptive and inferential, was employed to analyze this Excel workbook data set. This analysis was done by writing code in Python, with further visualizations from it using Tableau. From this we can compute correlations, probabilities, and other related metrics to predict cancellation rates for the future.

Overview of Statistical Analysis of the Data Set

Descriptive Statistics

- Describe each variable's features in Python (i.e. mean, standard deviation, quartiles, variable type)
- Convert or filter data as necessary and express visually using boxplots
- Barplots and additional boxplots comparing significant variables to cancellation (the dependent variable) in Tableau

Inferential Statistics

- Heatmapping of all variables to calculate correlation strengths (R2 values)
- OLS regression analysis to determine correlation of cancellation against the other variables
- Translating these results into a user-friendly practical framework to calculate cancellation probabilities

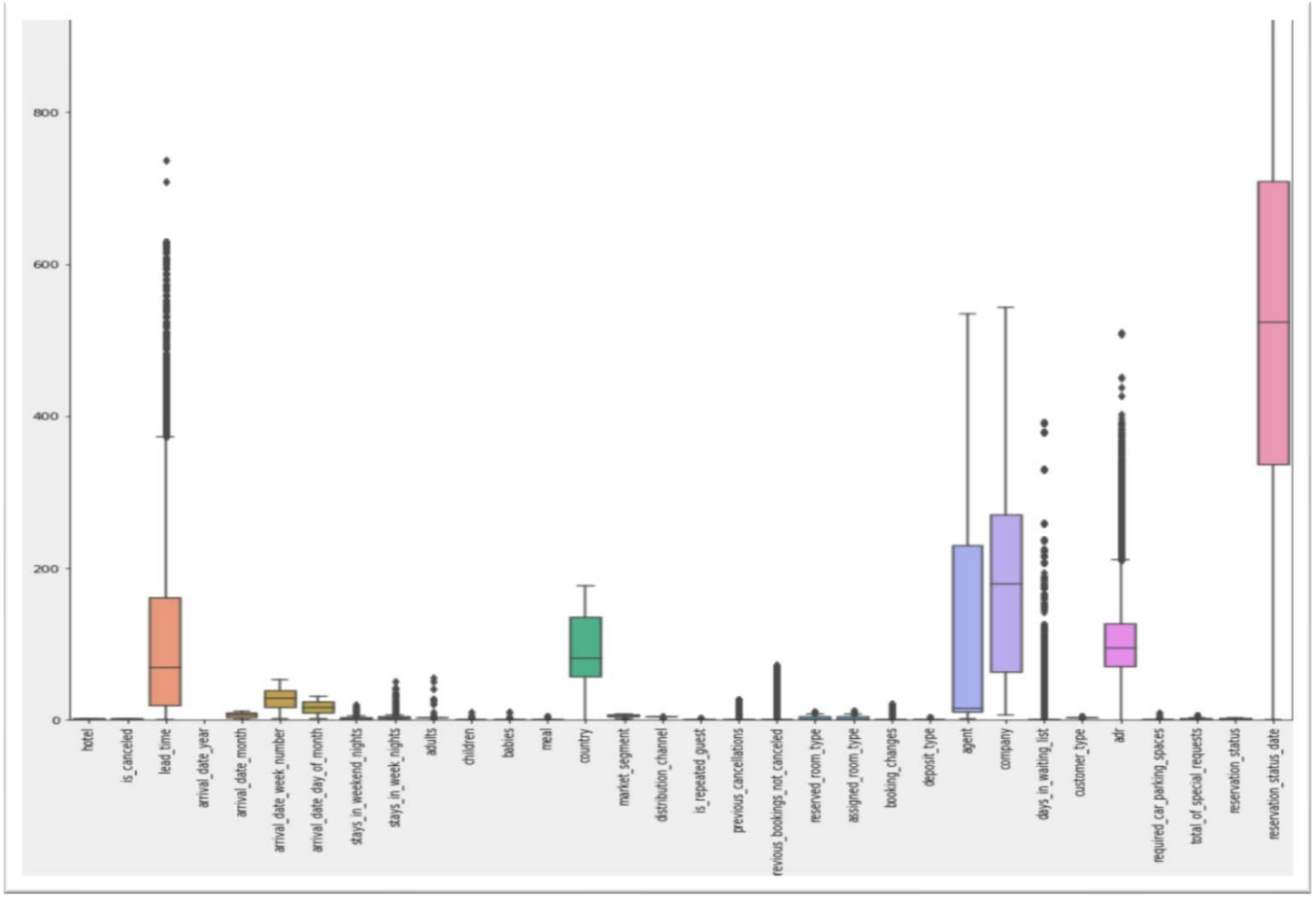
Description of Data

- 119,310 rows of data (unique reservations)
- 32 columns (variables) of data per reservation
- Several variable types, both numeric (integers, floats) and categorical (strings), convert as necessary

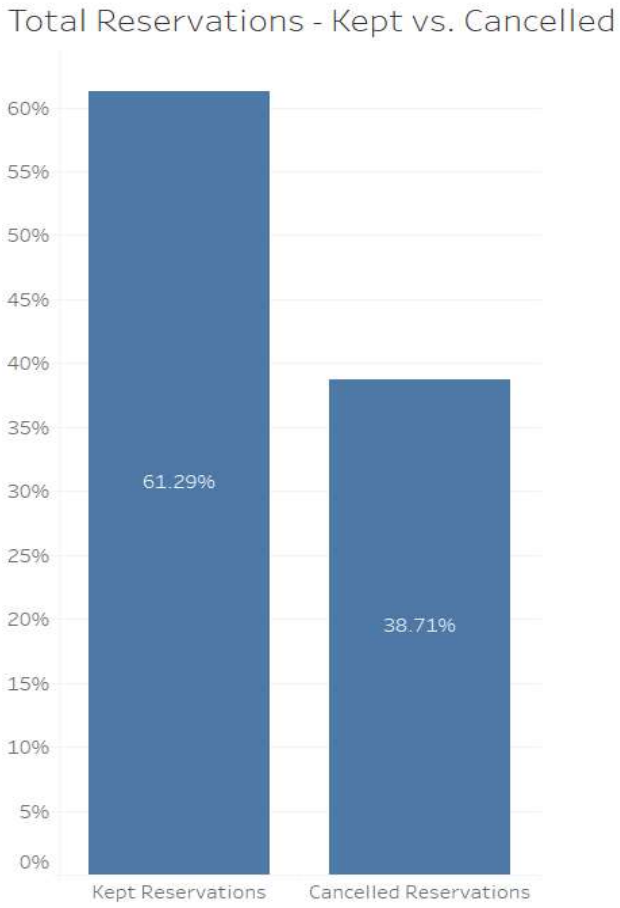
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
hotel	119390	2	City Hotel	79330	NaN	NaN	NaN	NaN	NaN	NaN	NaN
is_canceled	119390	NaN	NaN	NaN	0.370416	0.482918	0	0	0	1	1
lead_time	119390	NaN	NaN	NaN	104.011	106.863	0	18	69	160	737
arrival_date_year	119390	NaN	NaN	NaN	2016.16	0.707476	2015	2016	2016	2017	2017
arrival_date_month	119390	12	August	13877	NaN	NaN	NaN	NaN	NaN	NaN	NaN

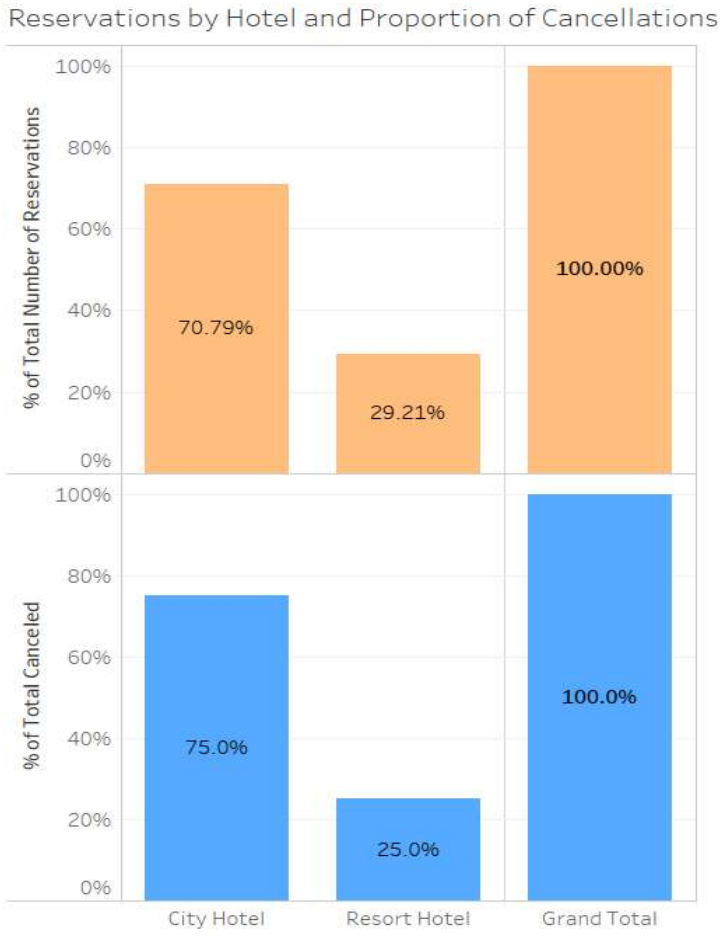
Initial boxplot of all data shows need to refine it:

- Large number of upper quartile outliers skew the ability to derive meaningful analysis – those outliers to be removed (note: 8 of 32 variables are null and removed b/c all of their values are such outliers)
- Wide scale between variables makes visualizing all to once difficult
- View cancellations ("is_canceled") as dependent variable and view others against it. About a quarter of variables stand out visually, for further review in Tableau



Descriptive Analysis of Cancellations and Hotel Types

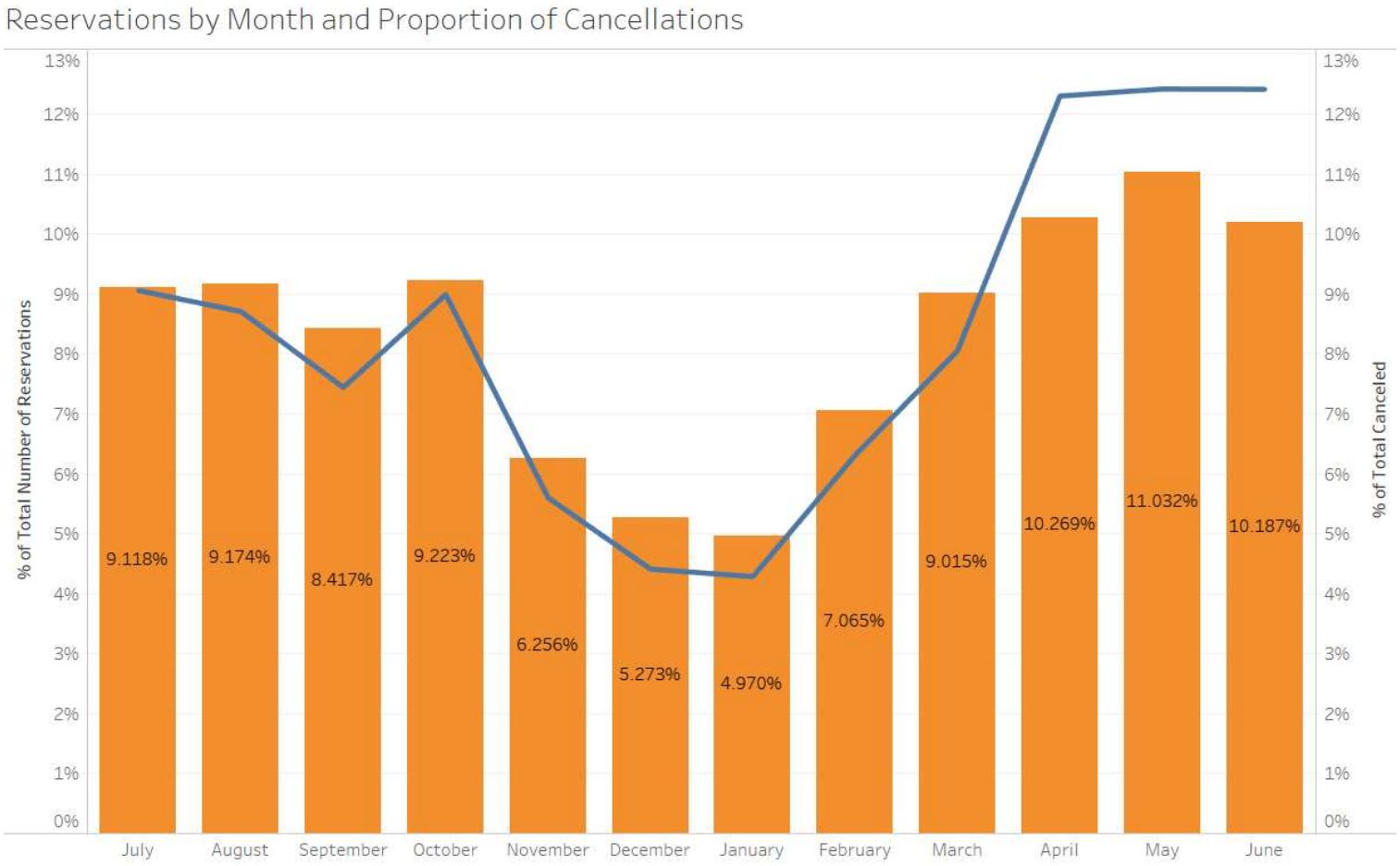




Insights

- **Over one-third of all guests cancelled their reservation**
- **Over 2/3 of all reservations were made at the City Hotel vs. the Resort Hotel, although a slightly higher proportion of cancellations occurred at the City Hotel (3/4 of total).**

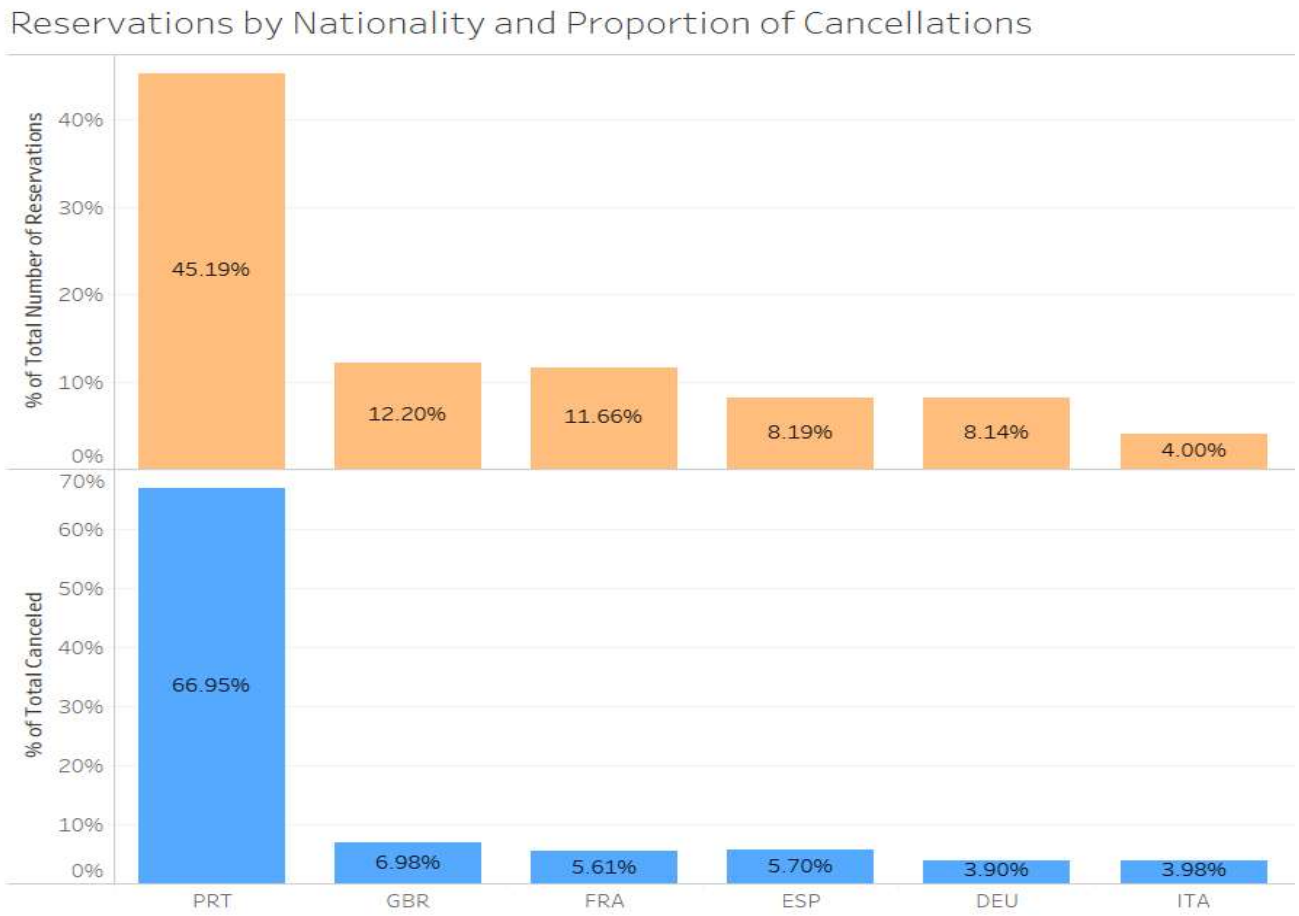
Descriptive Analysis of Seasonality



Insights

- **Reservations follow a seasonal pattern common in Portugal.**
- **Cancellations appear largely uncorrelated to month and are proportionately a product of the number of total reservations made. (Slight variation April-June)**

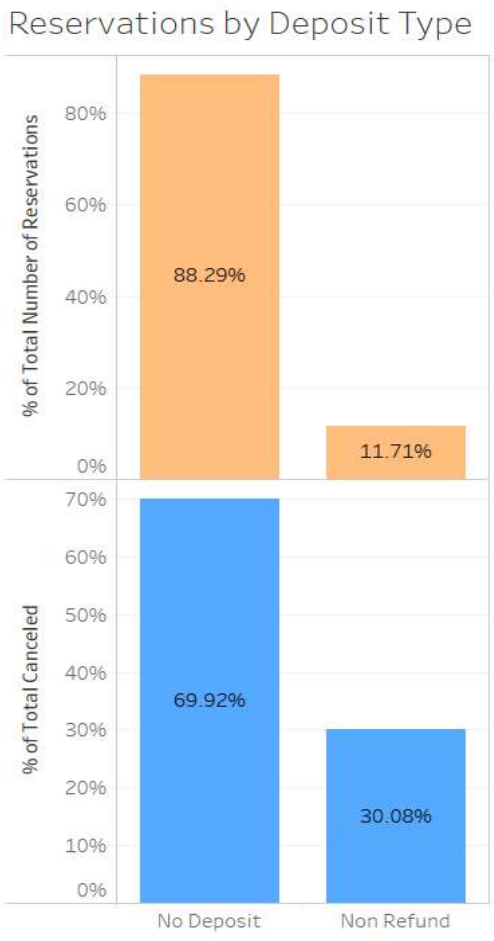
Descriptive Analysis of Guests' Nationality

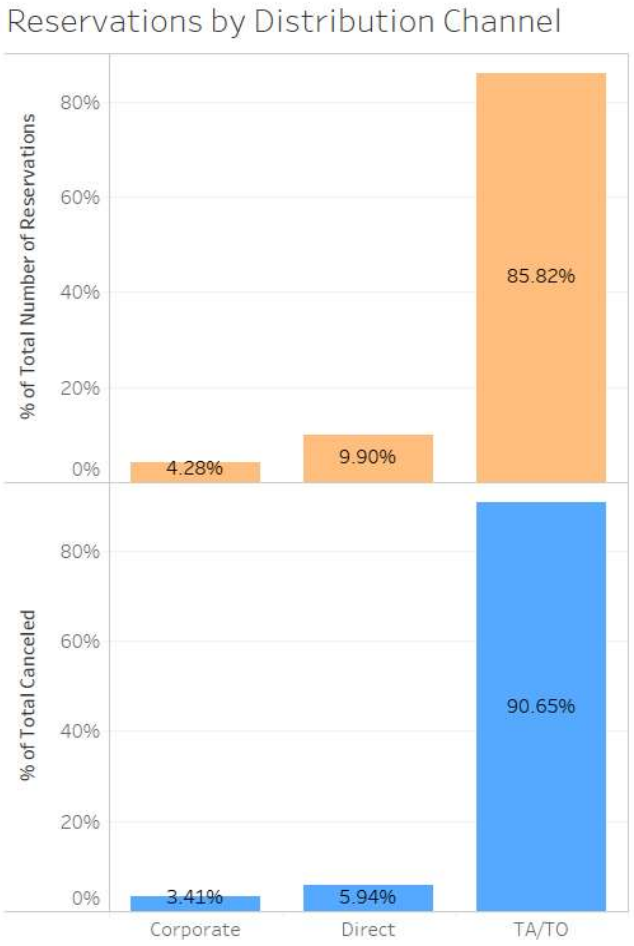


Insights

- **Almost half of all guests are local to Portugal, yet they make up a large majority of cancellations (2/3).**
- **Over 90% of guests are from Europe. Every other nationality contributes a disproportionately small share of cancellations.**
- **Note – guests come from over 150 countries, but only top few shown here.**

Descriptive Analysis of Distributions and Deposits

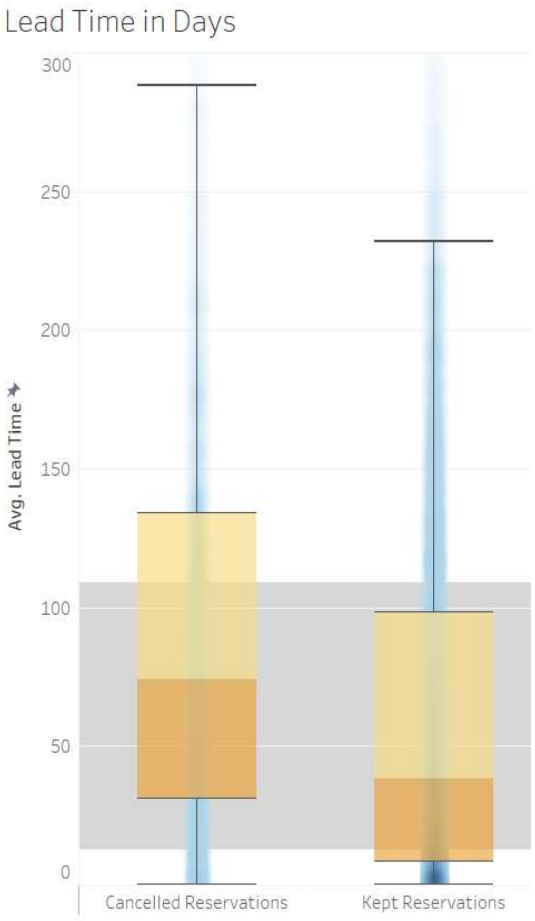


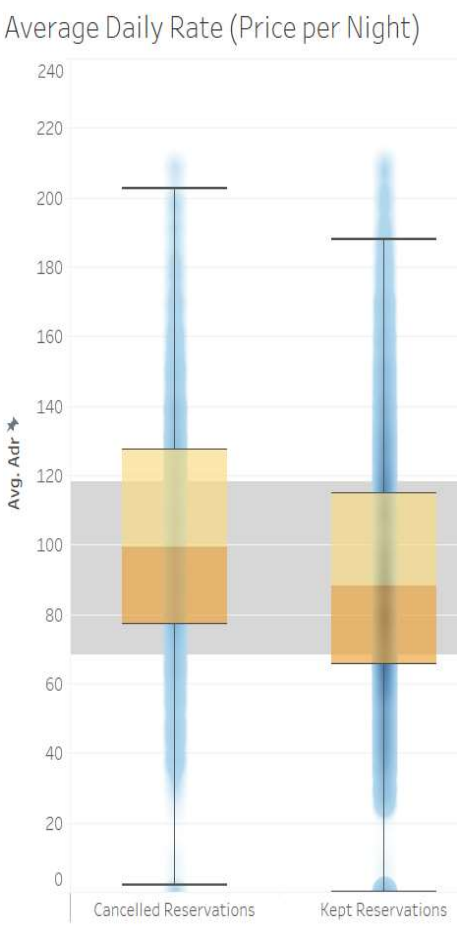


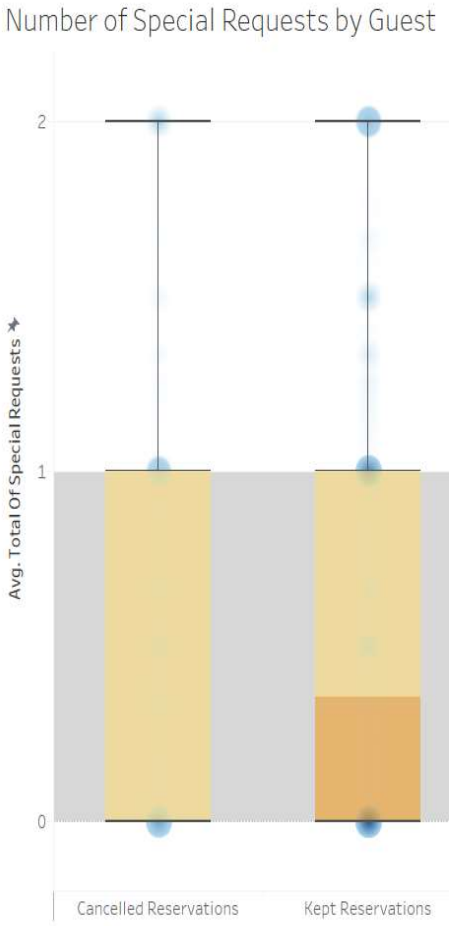
Insights

- **Counterintuitively, those who put down a deposit are more likely to cancel. Less than 1/8 of guests placed a deposit, but they account for almost 1/3 of cancellations.**
- **Most guests (over 85%) used an agent or operator to reserve, which is common in Europe. This majority was slightly more likely to cancel than those who booked directly or went through their company.**

Descriptive Analysis of Other Variables



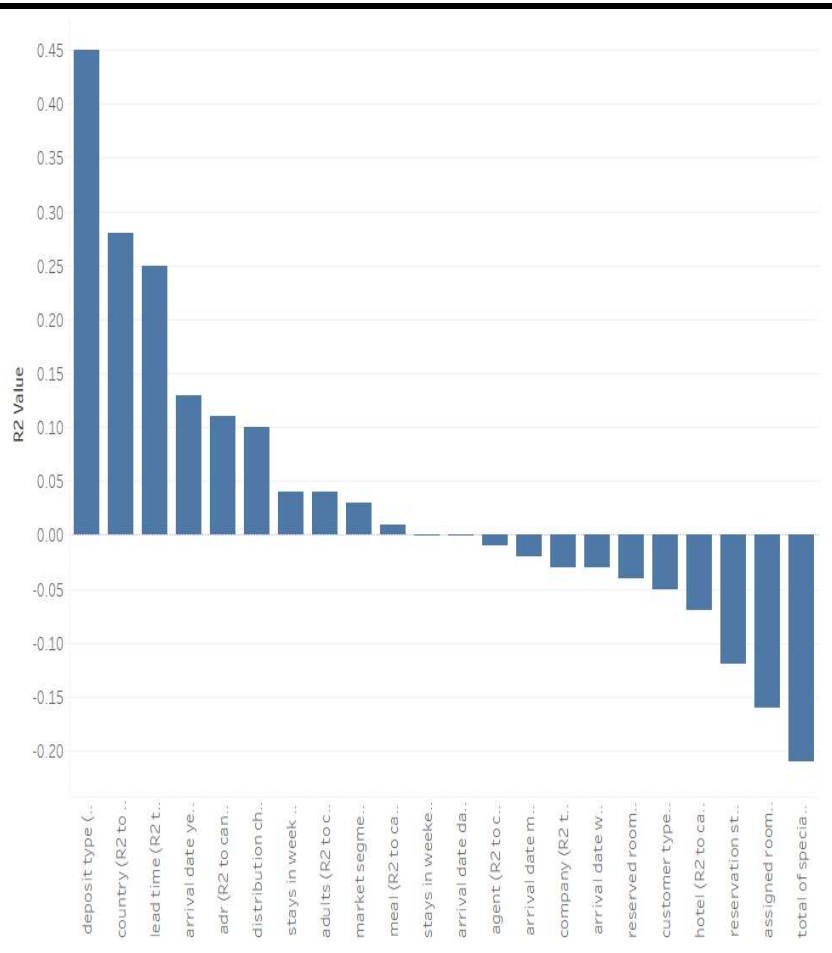




Insights

1. **Lead time – longer lead times result in a higher cancellation rates. The median lead time of those cancelled is nearly double that vs. kept (74 days vs. 38 days), although IQR's are similar (103 vs 90).**
2. **ADR – price drives a higher cancellation rate. The median cancelled reservation is 11 euro higher.**
3. **Special requests – more requests are made by those who don't cancel.**

Inferential Analysis – Correlations between All Variables and Cancellation



Insights

- Inferential statistical analysis here between all variables supports the observations noted in the previous slides.
- R2 values in heatmap between cancellations and other variables show strongest positive correlation with deposits, country (nationality), and lead time; strong negative correlation with total special requests and assigned room type.
- (Note – strongest negatively correlated variable, reservation status, is filtered out of regression analysis because it is an input directly associated with cancelled reservations.)
- Most variables studied are statistically insignificant (R2 below +/- 0.1).

Inferential Analysis – OLS Regression

Regression – All Variables

OLS Regression Results

Dep. Variable:	dependentVar	R-squared:	0.261
Model:	OLS	Adj. R-squared:	-0.478
Method:	Least Squares	F-statistic:	0.3527
Date:	Thu, 02 Apr 2020	Prob (F-statistic):	0.991
Time:	14:24:38	Log-Likelihood:	29.101
No. Observations:	45	AIC:	-12.20
Df Residuals:	22	BIC:	29.35
Df Model:	22		
Covariance Type:	nonrobust		

Regression – Top 7 Correlated Only

OLS Regression Results

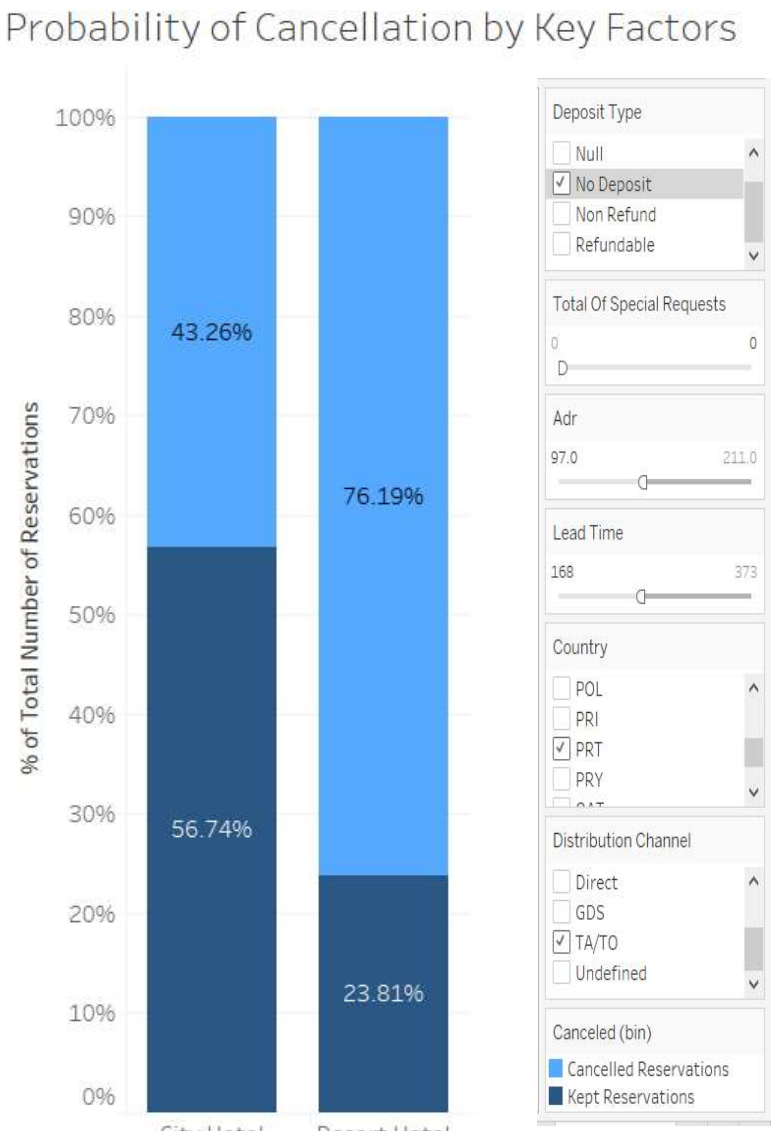
Dep. Variable:	dependentVar	R-squared:	0.266
Model:	OLS	Adj. R-squared:	0.266
Method:	Least Squares	F-statistic:	3520.
Date:	Thu, 02 Apr 2020	Prob (F-statistic):	0.00
Time:	14:29:58	Log-Likelihood:	-37015.
No. Observations:	67930	AIC:	7.405e+04
Df Residuals:	67922	BIC:	7.412e+04
Df Model:	7		
Covariance Type:	nonrobust		

- Deposit Type
- Country (Nationality)
- Lead Time
- ADR (Price)
- Hotel Type
- Distribution Channel
- Special Requests

Insights

- R2 of all variables to cancellation is 0.261 (left). Interestingly, the R2 of the top 7 variables is very similar, 0.266 (right). In predicting human behavior like this, low R2 values are expected because such behavior is inherently hard to predict and not surprising.
- Telling though, are other significant predictor measures, namely the Probability of the F-statistic; on the right it is 0.00 (much higher on the left w/all variables included). This mean there is 0% chance these 7 independent variables randomly affect our dependent variable (cancellation). There is sufficient evidence within the data to confirm these 7 affect the outcome.
- The coefficients (not shown) of those seven on the right can be used to write a multivariate, 7 variable equation that will calculate any given guest's probability of cancellation.
- This can also be performed through an interactive application like Tableau, as illustrated on the next slide.

Applied Use of Regression Analysis – Determining the Real Odds of Cancellation



Insights

- **Using the 7 most correlated pieces of information submitted by each guest determines the probability of cancellation.**

Example shown:

- **Portuguese guest**
- **Price of 97 euro per night**
- **No deposit**
- **Lead time of 168 days**
- **Booked through an agent (distribution channel)**

- **This guest has a 43% chance of cancelling at a City Hotel.**

- **The guest has a 76% chance of cancelling at a Resort Hotel.**

Conclusions – Practical Use for the Hospitality Manager

Using guest profiles as such, cancellation odds for each guest are predictable. Taken altogether, determining cancellation probability for an entire hotel for future dates becomes feasible.

Example:

- 80 guests have booked rooms at the city hotel
- Guests' combined cancellation probability is 40%.
- Hotel manager can anticipate that day only 48 guests arriving: ((80 – (40% x 80)) = 48).

Other uses to enhance hotel performance:

- Accurate revenue forecasting.
- Overbooking to maximize revenue. Especially during high season a hotel can confidently overbook to a certain point, based on cancellation odds, to maintain full occupancy.
- Operations - scale staffing, services, and supplies based on the anticipated number of guests.
- Track hotel's outreach to potential guests (i.e. deposits, special requests, how far in advance bookings are accepted).