Rental Prices- Explanatory Data Analysis **Investigation Overview** A dataset from a A housing rental company includes details about each property rented, Number of bedrooms, Number of bathrooms, as well as the price charged per night is provided. Data analysis of the the given data is done to answer the following questions: 1. What are main factors affect rental price?. 2. Do number of bathrooms has significant effect on price?. 3. Are there certain property or room types that have higher rental prices?. # Importing packages import numpy as np import pandas as pd import matplotlib.pyplot as plt import seaborn as sns %matplotlib inline # Loading data and printing out a few lines. df = pd.read csv('rentals cleaned csv', index col=0) Distribution of Rental Prices Price distribution is skewed, most pooular prices are under 20000\$ In [3]: # # univariate plot of price wirh setting Xlimit to get rid of outliers effect bins = np.arange(0, df['price'].max()+2000, 2000) plt.hist(df.price, bins = bins) plt.xlabel('Price (USD)') plt.ylabel('Frequncy') plt.title('Price distribution') plt.xlim((0,150000)); Price distribution 800 700 600 500 Frequncy 400 300 200 100 0 20000 40000 60000 80000 100000 120000 140000 Distribution of Bedrooms, Bathrooms and Minimum nights Most popular rentals have only one to two bedrooms, the same for bathrooms, We see that most of rentals are month rentals, follwoed by few 2 to 3 days rentals. # univariate plot of bathrooms In [4]: bins = np.arange(-.5, df['bathrooms'].max()+1, 1) plt.hist(data = df, x ='bathrooms', bins = bins, rwidth = 0.7) plt.xticks(np.arange(0, 14+1, 1)) plt.xlabel('No. of bathrooms') plt.title('Nmber of bathrooms distribution') plt.ylabel('count'); Nmber of bathrooms distribution 5000 4000 3000 2000 1000 0 No. of bathrooms # univariate plot of bedrooms bins = np.arange(-.5, df['bedrooms'].max()+1, 1)plt.hist(data = df, x = 'bedrooms', bins = bins, rwidth = 0.7) plt.xticks(np.arange(0, 14+1, 1)); plt.xlabel('No. of bedrooms') plt.title('Nmber of bedrooms distribution') plt.ylabel('count'); Nmber of bedrooms distribution 5000 4000 3000 2000 1000 9 10 11 12 13 14 8 2 6 No. of bedrooms # univariate plot of minimum nights df.minimum nights.value counts().head().plot(kind='bar') plt.xlabel('minimum nights') plt.ylabel('count') plt.title('Minimum nights distribution') plt.xticks(rotation=0); Minimum_nights distribution 3000 2500 2000 1500 1000 500 3 2 1 minimum_nights Price vs. Bedrooms and Bathrooms Price increases with the increase of the Number of bedrooms, bath rooms have low effect on price. # Define a function plot line heat() def plot scatter heat(df, price, column): """ Shows the relation between price and the other numerical column by returning a heat map and a scatterpl plt.figure(figsize = [18, 6]) plt.suptitle('relation between price and ' + column); # PLOT ON LEFT plt.subplot(1, 2, 1) sns.scatterplot(data = df, x = column, y = price, ci=False) plt.xlabel('Number of '+ column) plt.ylabel('Price') # PLOT ON RIGHT plt.subplot(1, 2, 2) plt.hist2d(data = df, x = column, y = price, cmin=0.5, cmap='viridis r') plt.colorbar() plt.xlabel('Number of '+ column) plt.ylabel('Price (USD)') # The relation between price and bathrooms plot_scatter_heat(df, 'price', 'bathrooms') relation between price and bathrooms 1.0 5000 800000 0.8 4000 600000 0.6 Price (USD) 3000 0.4 400000 2000 0.2 200000 1000 0.0 10 Number of bathrooms Number of bathrooms # The relation between price and bedrooms plot scatter heat(df, 'price', 'bedrooms') relation between price and bedrooms 1.0 5000 800000 0.8 4000 600000 0.6 3000 400000 2000 0.2 200000 1000 0.0 10 Number of bedrooms Number of bedrooms **Price and Room Types and Property Types** There is a relationship between price property types of Resort, Villa, Earth house, and the room type of Entire home/apt as they have the higest mean prices, The two combinations of Entire home/apt and villa has higest prices. # The relation beween room type and mean price base_color = sns.color_palette()[0] sns.barplot(data =df, x= 'room_type', y= 'price', color= base_color) plt.ylabel('mean price (USD)') plt.title('relation beween room type and mean price') plt.rcParams['figure.figsize'] = 8,6; relation beween room type and mean price 25000 20000 mean price (USD) 15000 10000 5000 0 Entire home/apt Shared room Private room Hotel room room_type In [19]: # The relation beween property type and price base_color = sns.color_palette()[0] sns.barplot(data =df, x= 'property_type', y= 'price', color= base_color, ci=False) plt.xticks(rotation=90) plt.ylabel('mean price (USD)') plt.title('relation beween property type and mean price') plt.rcParams['figure.figsize'] = 12,10; relation beween property type and mean price 50000 40000 30000 mean price (USD) 20000 10000 House Guest suite Cottage Hostel Bed and breakfast Hotel Boutique hotel Other Castle Cabin Townhouse Serviced apartment Tiny house Resort Villa In-law Earth house Dome house Condominium Ψ Guesthouse Bungalow Aparthotel Camper/RV property_type # Relation between room type, property type and mean price df.groupby(['room_type','property_type'])['price'].mean().plot(kind= 'bar') plt.ylabel('mean price (USD)') plt.title('relation beween room and property types and mean price') plt.rcParams['figure.figsize'] = 15,6; relation beween room and property types and mean price 250000 200000 mean price (USD) 150000 100000 50000 (Entire home/apt, Condominium) (Entire home/apt, Cottage) (Entire home/apt, Dome house) (Entire home/apt, Earth house) (Entire home/apt, Guesthouse) (Entire home/apt, Hotel) (Entire home/apt, House) (Entire home/apt, In-law) (Entire home/apt, Loft) (Entire home/apt, Other) (Entire home/apt, Serviced apartment) (Entire home/apt, Tiny house) (Entire home/apt, Townhouse) (Entire home/apt, Villa) (Hotel room, Aparthotel) (Private room, Aparthotel) (Private room, Apartment) (Private room, Cabin) (Private room, Earth house) (Private room, Guest suite) (Private room, Hotel) Private room, House) (Private room, Serviced apartment) (Private room, Townhouse) (Shared room, Bed and breakfast) (Entire home/apt, Guest suite) (Hotel room, Resort) (Private room, Cottage) (Shared room, Serviced apartment) room_type,property_type Price and Number of Bath rooms and Bedrooms by Room and Property Types From the multivariate exploration we know that the reason was due to different categories types of room and property, The Entire home/apt was the room type was the only one to increase in price when bathroom and bedroom numbers increases. # Define a function plot line category() def plot scatter category(df, price, column, categroy): """ Shows the relation between price and the other numerical column according to a specefic category type plt.figure(figsize = [18, 6]) plt.title('relation between price and ' + column + ' by ' + categroy); sns.scatterplot(data = df, x = column, y = price, hue=categroy ,ci=False) plt.xlabel('Number of '+ column) plt.ylabel('Price (USD)') plot scatter category(df, 'price', 'bathrooms', 'room type') relation between price and bathrooms by room_type 1.0 room_type Entire home/apt Private room Shared room 0.8 Hotel room 0.2 0.0 12 plot scatter category(df, 'price', 'bedrooms', 'room type') relation between price and bedrooms by room_type 1.0 Entire home/apt Private room Shared room 0.8 0.2 10 Number of bedrooms plot scatter category(df, 'price', 'minimum nights', 'room type') relation between price and minimum_nights by room_type 1.0 Entire home/apt Private room Shared room 0.8 0.2 100 200 Number of minimum_nights **Regression Model** We finished with some Regression to predict the price. the top important features were bedrooms, minimum nights, bathrooms_bderooms, and property type of twon house. In [46]: dummies = pd.get dummies(df, prefix sep=' ') dummies['bathrooms bedrooms'] = dummies['bathrooms'] * dummies['bedrooms'] X = dummies.drop(['id','latitude' ,'longitude','bathrooms','price','property type Aparthotel','room type Shared = df[['price']] # Import train test split from sklearn.model_selection import train_test_split # Split into train and test sets X_train, X_test, y_train, y_test = train_test_split(X, test size=.33, random state=42) # Import MinMaxScaler from sklearn.preprocessing import MinMaxScaler # Instantiate MinMaxScaler and use it to rescale X train and X test scaler = MinMaxScaler(feature range=(0, 1)) rescaledX_train = scaler.fit_transform(X_train) rescaledX test = scaler.transform(X test) # Import Regressors from sklearn.linear model import LinearRegression, Ridge, Lasso, ElasticNet # Import mean_squared_error as MSE from sklearn.metrics import mean squared error as MSE # Set seed for reproducibility SEED = 1# Instantiate regressors In [47]: reg = LinearRegression() #fit reg to the training set reg.fit(rescaledX train, y train) # Predict the labels of the test set y pred = reg.predict(rescaledX test) # Evaluate the test-set RMSE of reg on the test set print('{:s} : {:.3f}'.format('linear regression MSE', MSE(y test, y pred)**.5)) linear regression MSE : 27998.986 In [48]: # Create a plot of features importances using random forest from sklearn.ensemble import RandomForestRegressor rf = RandomForestRegressor(random state=SEED) rf.fit(rescaledX_train, y_train) <ipython-input-48-2bc682d68965>:4: DataConversionWarning: A column-vector y was passed when a 1d array was expe cted. Please change the shape of y to (n_samples,), for example using ravel(). rf.fit(rescaledX_train, y_train) Out[48]: RandomForestRegressor(random_state=1) In [78]: # Create a pd.Series of features importances importances_rf = pd.Series(rf.feature_importances_, index = X.columns) # Sort importances rf sorted_importances_rf = importances_rf.sort_values()[13:] # Make a horizontal bar plot sorted_importances_rf.plot(kind='barh', color='lightgreen'); plt.show() plt.rcParams['figure.figsize'] = 12,6 bedrooms minimum_nights bathrooms_bedrooms property_type_Townhouse property_type_Condominium property_type_House room_type_Entire home/apt property type Apartment room type Private room property_type_Hotel property_type_Boutique hotel property_type_Serviced apartment property_type_Guest suite property_type_Resort room_type_Hotel room property_type_Villa property_type_Loft property_type_Bed and breakfast 0.05 0.15 conclusion From previous explanatory data analysis data analysis we can conclude the follwoing conclusion: 1. Bedrooms have the higest effect on rental peice, followed by the minimum nights. 2. There is a linear relation between price, and numerical features except for the bathrooms that is replaced by the interaction term of bathrooms multiplied by bedrooms, actualy we don't need bathrooms in our model, also there is no multimulticollinearity. 3. Price distribution is skewed, most common prices are under 20000\$. 4. Most popular rentals have only one to two bathrooms. 5. Most popular rentals have only one to two bedrooms. 6. Most of rentals are month rentals, follwood by few 2 to 3 days rentals. 7. Most common room and property type are the apartment type for both the room and the property. 8. There is a relationship between price and property types of Resort, Villa, Earth house, and the room type of Entire home/apt as they have the higest mean prices. 9. The two combinations of Entire home/apt and villa have the higest prices. 10. The Entire home/apt was the room type was the only one to increase in price when bathroom and bedroom numbers increases. 11. The top important features were bedrooms, minimum nights, bath_bed rooms ,and property type of twon house. jupyter nbconvert Rental_prices_explanatory_data_analysis.ipynb --to slides --post serve --no-input --no-prome!