08/03/2021

CURNEU MEDTECH INNOVATIONS PRIVATE LIMITED – SD03Q014

REPORT – 02

Problem Statement:

To understand and fit the suitable Machine Learning model for the given dataset "houses\_to\_rent.csv" and fetch best accuracy for the fitted model.

Abstract:

The main maxim is to find the total rent amount for houses in Brazil besides having property tax, fire insurance, house insurance, furnished home, and allowance of pets. Initially Exploratory Data Analysis will be done to explore the data in a deeper way and suitable model will be fitted after crucial analysis of each and every variable. Finally the statement must be concluded by predicting the house rent for the houses.

Dataset Description:

There are basically 13 attributes and around 10,000 data. It includes city, area, rooms, bathroom, parking spaces, number of floors, acceptance of pets, furnished homes or not, homeowner association, rent amount, property tax, fire insurance and finally the total amount in dollars.

EDA:

The graphs were plotted using these attributes.

* Initially heat map was plotted to find the correlation between the attributes.
* Next a pie-chart was plotted to find the most popular city among Brazilians which resulted in "Sao Paulo".
* Another heat map was plotted to find the correlation between the acceptances of pets and furnished/not houses. Since these attributes were categorical the relation between them was calculated here.
* A scatter plot was plotted to find the correlation between property tax, furnished home, and property tax, acceptances of pets which showed that they correlate much which states that the tax directly depends on the mentioned two attributes.
* Some set of box plots were plotted, "Boxplot of Rent amount by parking spaces", "Boxplot of Rent amount by rooms", "Boxplot of Rent amount by bathroom", "Boxplot of Rent amount by city". They showed less outliers and much correlation between the attributes.
* Another set of box plots were plotted, "Boxplot of Total amount by fire insurance", "Boxplot of Total amount by property tax", " Boxplot of Total amount by floor", "Boxplot of Total amount by rent amount". They showed relatively more outliers.
* A count plot was plotted between the attributes "Total", "Animal" and "Total", "Furniture".
* Set of bar plots were plotted for "Area", "Rooms", "Bathroom", "Floor" to group the values.
* Various bar plots were plotted as "Average area in a city", "Average number of rooms of a house in each city", "Maximum number of rooms of a house in each city" by grouping the values under mean function for easy visualizations.
* Pairplot was plotted for train set which showed correlatory effects between all the attributes.

Code:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

sns.set(color\_codes=True)

%matplotlib inline

dt = pd.read\_csv(r'houses\_to\_rent.csv')

dt.head(10)

dt["floor"].replace("-", 0, inplace=True)

dt["floor"] = dt["floor"].astype("int")

dt.info()

dt.shape

dt.describe()

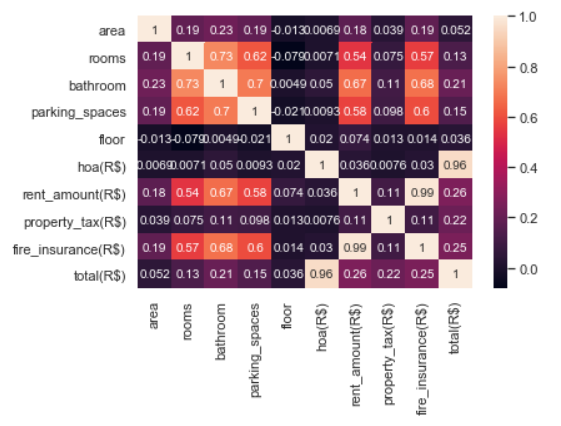
dt.isnull().values.any()

cor = dt.corr()

print(cor)

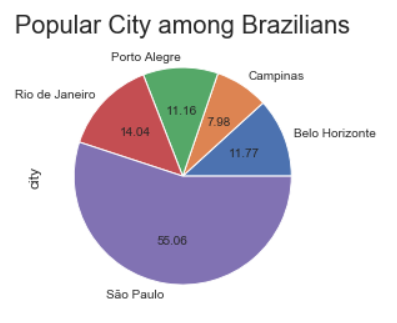
sns.heatmap(cor,annot=True)

plt.show()



dt.groupby(['city'])['city'].aggregate(lambda x: x.count()/ 10692).plot(kind='pie',autopct='%.2f',fontsize=10)

plt.title("Popular City among Brazilians",fontsize=20)



sns.heatmap(dt.groupby(['animal','furniture']).size().unstack(), annot=True, fmt="d")

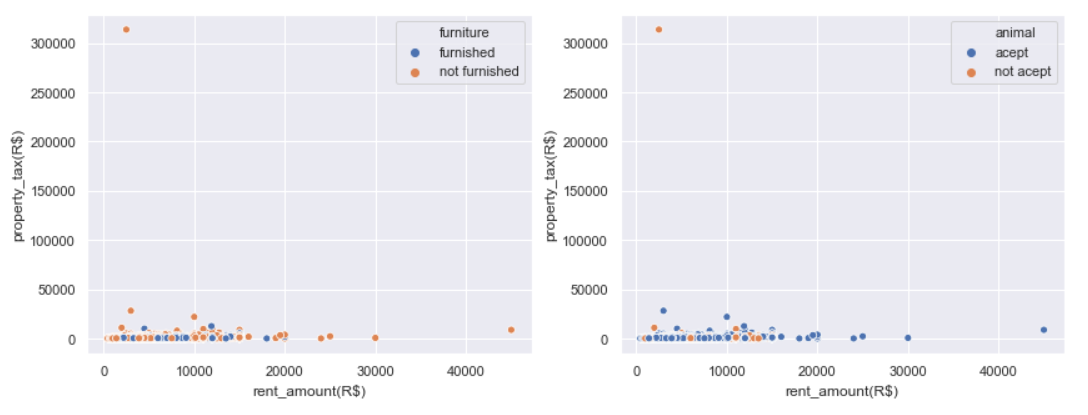
plt.title("Animal & Furniture",fontsize=20)



fig, axes = plt.subplots(1,2, figsize=(14,5))

sns.scatterplot(data=dt, y='property\_tax(R$)', x='rent\_amount(R$)', hue="furniture", ax=axes[0])

sns.scatterplot(data=dt, y='property\_tax(R$)', x='rent\_amount(R$)', hue="animal", ax=axes[1])



list\_of\_metrics = [['parking\_spaces','rooms'], ['bathroom', 'city']]

def box\_plot\_func1(ax, metric):

dt.boxplot(column='rent\_amount(R$)', by=f'{metric}', fontsize=10, ax=ax);

ax.set\_title("Boxplot of Rent amount by " + f"{metric}")

ax.set\_xlabel(f"{metric}")

ax.set\_ylabel('Rent amount')

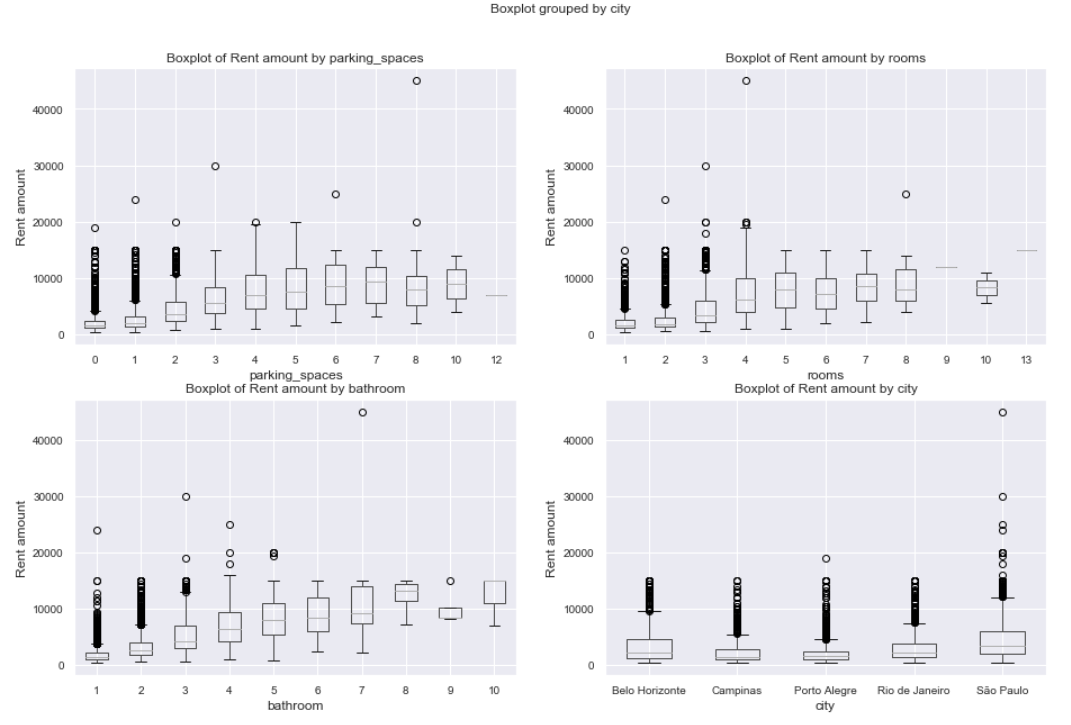
fig, axes = plt.subplots(2,2, figsize=(15,10))

box\_plot\_func1(axes[0,0], list\_of\_metrics[0][0])

box\_plot\_func1(axes[0,1], list\_of\_metrics[0][1])

box\_plot\_func1(axes[1,0], list\_of\_metrics[1][0])

box\_plot\_func1(axes[1,1], list\_of\_metrics[1][1])



list\_of\_metrics = [['fire\_insurance(R$)','property\_tax(R$)'], ['floor', 'rent\_amount(R$)']]

def box\_plot\_func2(ax, metric):

dt.boxplot(column='total(R$)', by=f'{metric}', fontsize=10, ax=ax);

ax.set\_title("Boxplot of Total amount by " + f"{metric}")

ax.set\_xlabel(f"{metric}")

ax.set\_ylabel('Total amount')

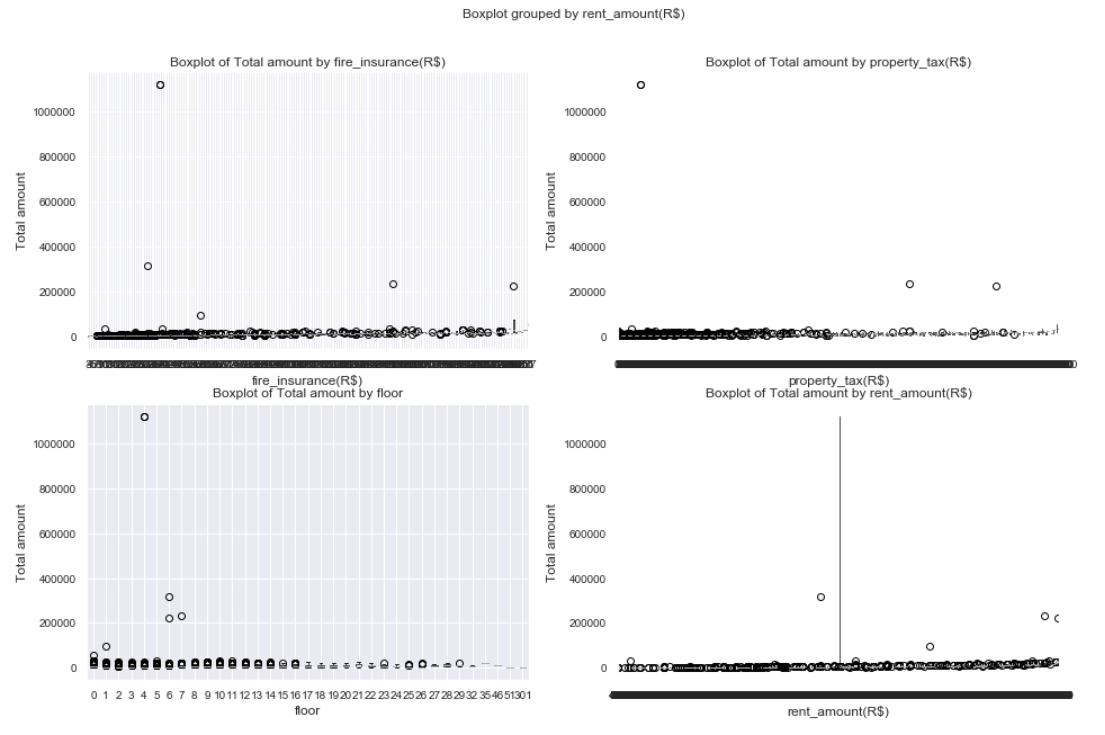
fig, axes = plt.subplots(2,2, figsize=(15,10))

box\_plot\_func2(axes[0,0], list\_of\_metrics[0][0])

box\_plot\_func2(axes[0,1], list\_of\_metrics[0][1])

box\_plot\_func2(axes[1,0], list\_of\_metrics[1][0])

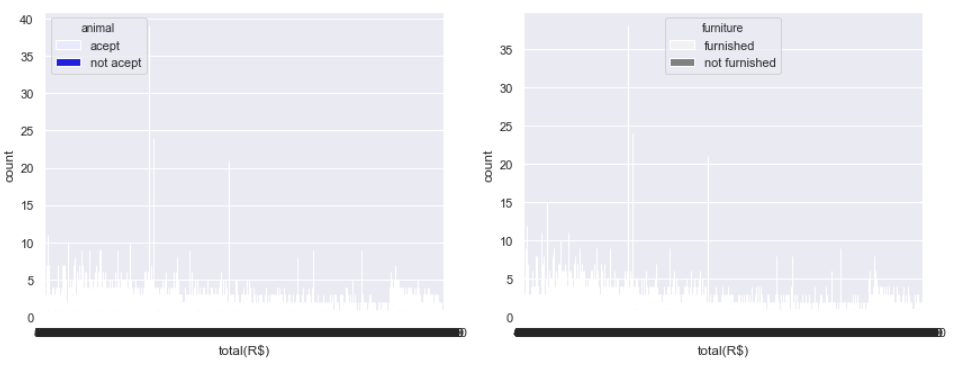
box\_plot\_func2(axes[1,1], list\_of\_metrics[1][1])



fig, axes = plt.subplots(1,2, figsize=(14,5))

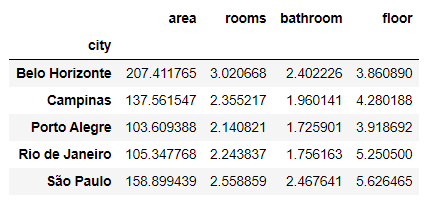
sns.countplot(x='total(R$)', hue='animal', data=dt, ax=axes[0], color="Blue");

sns.countplot(x='total(R$)', hue='furniture', data=dt, ax=axes[1], color="Grey");



dt\_group = dt.groupby(["city"])[["area","rooms","bathroom","floor"]].aggregate(lambda x: x.mean())

dt\_group



fig, ax = plt.subplots(1, figsize=(16, 4))

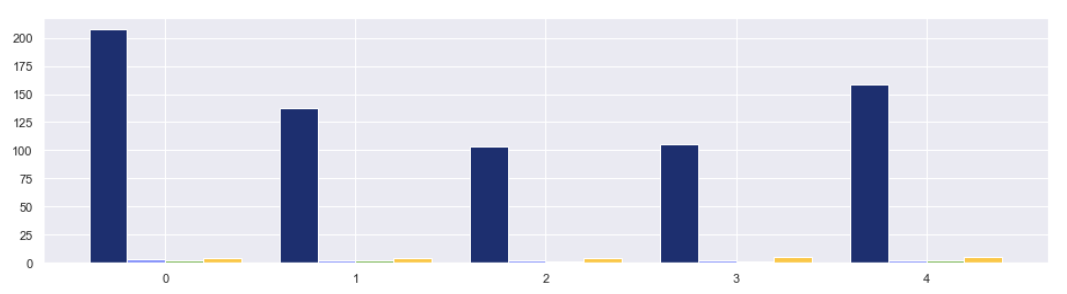
x = np.arange(0, len(dt\_group.index))

plt.bar(x - 0.3, dt\_group["area"], width = 0.2, color = '#1D2F6F')

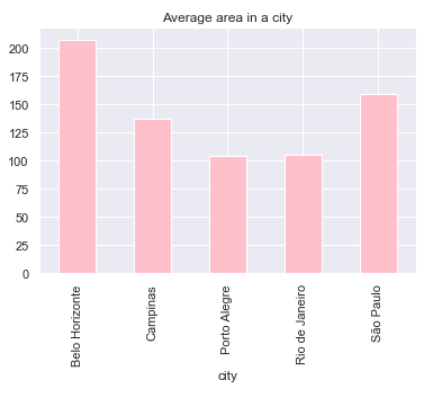
plt.bar(x - 0.1, dt\_group["rooms"], width = 0.2, color = '#8390FA')

plt.bar(x + 0.1, dt\_group["bathroom"], width = 0.2, color = '#6EAF46')

plt.bar(x + 0.3, dt\_group["floor"], width = 0.2, color = '#FAC748')

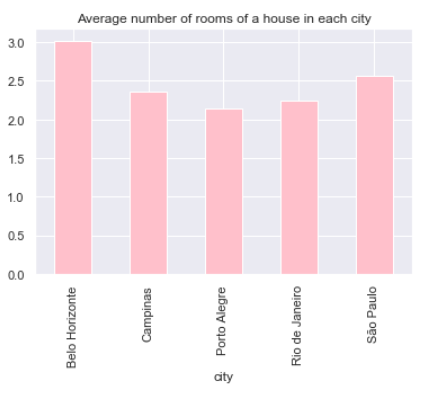


dt.groupby(['city'])['area'].aggregate(lambda x: x.mean()).plot(kind="bar",color="Pink")

plt.title("Average area in a city", fontsize=12)

dt.groupby(['city'])['rooms'].aggregate(lambda x: x.mean()).plot(kind='bar',color="Pink")

plt.title('Average number of rooms of a house in each city', fontsize=12)



dt.groupby(['city'])['rooms'].aggregate(lambda x: x.max()).plot(kind='barh',color="Red")

plt.title('Maximum number of rooms of a house in each city', fontsize=12)



dt.isna().sum()

from sklearn.model\_selection import train\_test\_split

del dt['total(R$)']

x, y = dt.drop('rent\_amount(R$)',axis=1), dt['rent\_amount(R$)']

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2)

print(f'x train shape: {x\_train.shape}, x test shape: {x\_test.shape}')



from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

x\_train.loc[:,'city'] = le.fit\_transform(x\_train['city'])

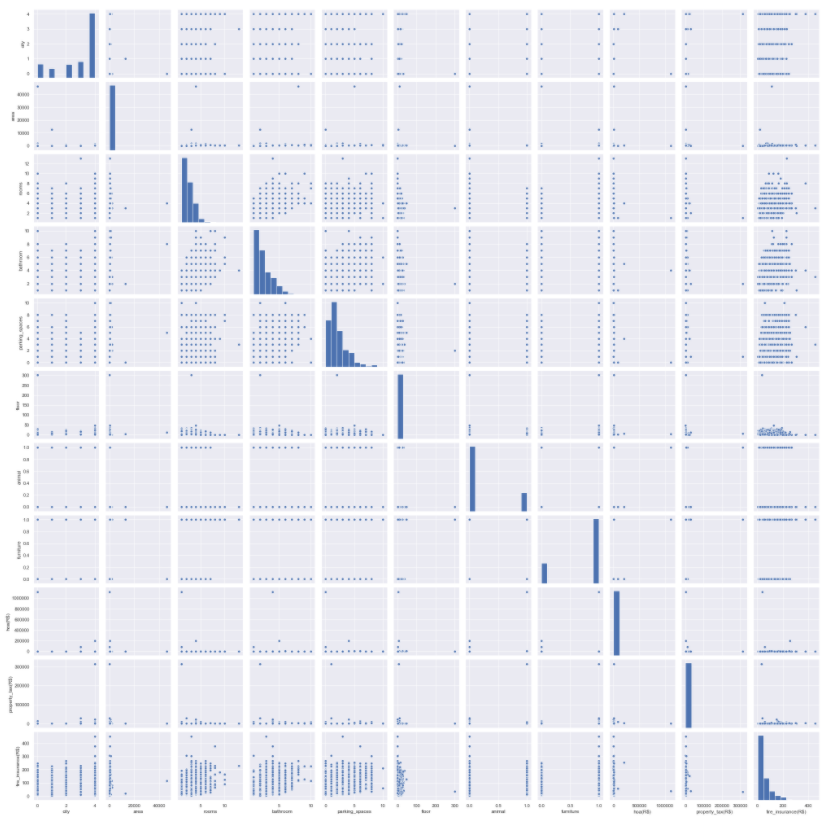
x\_test.loc[:,'city'] = le.transform(x\_test['city'])

x\_train.loc[:,'animal'] = le.fit\_transform(x\_train['animal'])

x\_test.loc[:,'animal'] = le.transform(x\_test['animal'])

x\_train.loc[:,'furniture'] = le.fit\_transform(x\_train['furniture'])

x\_test.loc[:,'furniture'] = le.transform(x\_test['furniture'])

sns.pairplot(x\_train)

from sklearn.model\_selection import train\_test\_split

from sklearn import model\_selection

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error

from sklearn.metrics import r2\_score

from math import sqrt

x = dt[["area","rooms","bathroom","parking\_spaces","hoa(R$)","property\_tax(R$)","fire\_insurance(R$)"]]

y = dt[["rent\_amount(R$)"]]

lr = LinearRegression()

lr.fit(x\_train, y\_train)

pred\_train\_lr = lr.predict(x\_train)

pred\_test\_lr = lr.predict(x\_test)

regressor = LinearRegression()

regressor.fit(x\_train,y\_train)

y\_pred = regressor.predict(x\_test)

print("RMSE and r-square for train set:")

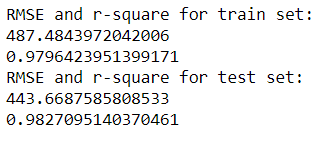
print(np.sqrt(mean\_squared\_error(y\_train,pred\_train\_lr)))

print(r2\_score(y\_train,pred\_train\_lr))

print("RMSE and r-square for test set:")

print(np.sqrt(mean\_squared\_error(y\_test,pred\_test\_lr)))

print(r2\_score(y\_test,pred\_test\_lr))



Conclusion:

From the graphs, it can be inferred that the house rent increases each year. The accuracy of train set and test set are 97% and 98% respectively. From the graphs, it can be assumed that the houses in Sao Paolo must be higher since it is the most popular city in Brazil.