Seazone Challenge Report

The present report aims to show what I have done to accomplish the “Seazone Challenge” in the best way I could, covering its difficulties, solutions found, and how I would improve it on the future.

Once I´ve created my repository, the first problem I encountered was that, the given data was to large (specifically the ‘Price_AV_Itapema.csv’ file) to be stored in my repository. The data had to be accessed by the codes on the repository, so I transformed the file into a zipped file and using Git LFS (a Git extension that enables the storage of larger files on the repository) so it could be upload and accessed. And the first change I would do to improve my solution would be to find or pay for another way to store and access the data, because uploading large files on to Github is not a good practice. Considering a professional situation these files would get larger and eventually they wouldn’t fit on a Git’s repository. A possible solution would be storing the data on the cloud (an Amazon S3 instance for example), a paid service would give more scalability for the whole situation, decrease the memory usage, reduce the script’s processing time and save me more time for focusing on the business’ solutions.

After configuring the repository (adding the requirements needed, specifying the folders, testing on a virtual environment and if it was all set) I actually started checking the data. The whole case is based on properties (houses and apartments), containing four datasets from Airbnb site and one from VivaReal. As soon as I’ve read the case’s description, I decided that using a regression machine learning method would be a good way of solving the problems presented, so I started the Data Wrangling process aiming on modeling the data in a way that it would be possible to use the best method. Each one of the Airbnb datasets covers different aspects, containing different paramethers, from the properties, so I needed to check if the datasets were encompassing the same group of properties and I found that they were. This way, it would be possible to join all the Airbnb data in just one dataset, containing the most critical information about the properties.

The first Data Wrangling script was done in order to create a modeled dataset from the Details_Data.csv file. As soon as I started checking the file, I noticed that it contained various registers – encompassing different dates - from the same property, and assumed that keeping just one register per property would be the best way to model its data. This dataset also had some columns that wouldn’t be used (that I excluded) and some composed by categorical paramethers, concerning amenities, house rules, among other information about the properties, so I vectorized this paramethers in order to model a dataset that could fit on a machine learning method and saved the modeled dataset.

The second Data Wrangling script was done in order to create a modeled dataset from the Prices_AV_Itapema.zip file, where I started dropping duplicates, excluding registers with null values on critical columns, and excluding some outliers. Even after these proccesses, in order to keep just on register per property, I summarized the price and the “availability data” (which were the most important for solving the presented problems), keeping the properties main price, available rate (A metric that I created that represents the percentage of registers on which the property was available), total revenue (calculated considering that if the property is not available on that moment, it's been rented), the minimum stay mode and the number of different dates that the property was listed and saved at the modeled dataset.

Once I had the two modeled datasets, I aggregated them with the 'Mesh_Ids_Data_Itapema.csv', which contains the most reliable information about Airbnb properties; latitude and longitude. I also added on it hosts' information, from 'Hosts_ids_Itapema.csv'. At this point I had transformed all four Airbnb datasets into one condensed dataset, modeled in a way that it could be fitted onto a machine learning method and having 0,02% of the raw data size. A last check at the modeled data was needed so I could correct some incongruencies, exclude duplicated columns, among other minor cleaning processes. The whole proccess of cleaning, wrangling, enriching and modeling the Airbnb raw data into a modeled dataset was necessary for many reasons, but in some cases it had a few disadvantages, for example, the modeled data desconsidered the information related to the dates, so if was needed to get some insights concerning the most rentable months the condensed dataset wouldn't be useful (and it would be modeled in a different way).

All the information from VivaReal site was contained on VivaReal_Itapema.csv file, but I used the same logic from the previous scripts, aiminig to create a modeled database. The first steps were excluding columns that wouldn't be used, dropping duplicates, dropping registers with null values that couldn't be treated, among others. Checking the price columns (the dataset encompasses properties for sale and rent) I noticed the presence of outliers, and rather than just removing them, I tried to treat them in another way, finding some ouliers among the outliers (and removing only these registers). The column that indicates the property's neighborhood information was also very important, but some registers had no values on it, so I've used some NLP methods in order to acquire the property's neighborhood information from the listing title. Some categorical columns also needed to be vectorized before saving the two modeled datasets, concerning both the "for rental" and "for sale" properties.

At this point, the DataWrangling proccess was finished, and I started to solve the proposed problems. For the first one - aiming to find the best propery profiles to invest on the city – I used the Airbnb_Modeled_Dataset.csv and its biggest challenge started on the definition of the concepts 'best' and 'profile'. I defined 'best investiments' based on its average price and availability rate (the metric that I'd created on the second Data Wangling script) and profile as the property type (House/Appartment/Loft). Considering I've excluded registers with more than one week as the minimum staying days. I noticed that the price of some property types as "Pousada", "Bangalô" and "Vila" was considerably higher for renting, and it could be a good investiment. Considering the most common types of property, such as: "Casa", "Apartamento", and "Condomínio", I found that there's almost no difference between their availability rates. Their average price are:

- "Casa": R$621/day,
- "Apartamento": R$676/day;
- "Condomínio": R$730/day.

.

I applied the Random Forest Regression Algorithm for the most commom types, such as: “Casa”, “Apartamento”, and “Condomínio” that could predict its price with 90% accuracy based on its features. The number of rooms (bathrooms, bedrooms) and number of guests are the features that have bigger impacts on the average price, and this kind of properties are probably the ‘best investments’.

The second problem was solved in a simpler way. I’ve defined location as the neighborhood set on the VivaReal_Itapema.csv file and checked that the properties on the “meia praia” and “andorinha” were most valuable.

For the third problem resolution I used the Modeled_Airbnb_Data.csv file, on which I excluded outliers and separated a list of useful paramethers. I’ve used the K-means Clustering model for aggregating the geolocation zones based on the properties proximities, and vectorized the column for applying the Random Forest Regression algorithm to check the more important features for increasing the properties’ values.

The fouth problem’s resolution was very similar, but considering only the properties which had the "Espaço inteiro: apartamento" property type. I’ve also used the K-means Clustering model for aggregating the geolocation zones with their revenue values. The data was divided in 6 clusters, with two of them having a considerably bigger revenue value than the rest. I’ve used the Random Forest Regression in order to find the most relevant features for increasing the revenue, and the insights were the same: “more rooms” = “more valuable price”. For calculating the builgs return over the year I’ve calculated the revenue daily value for properties on the two most valuable zones with a number of rooms bigger than the Data median.