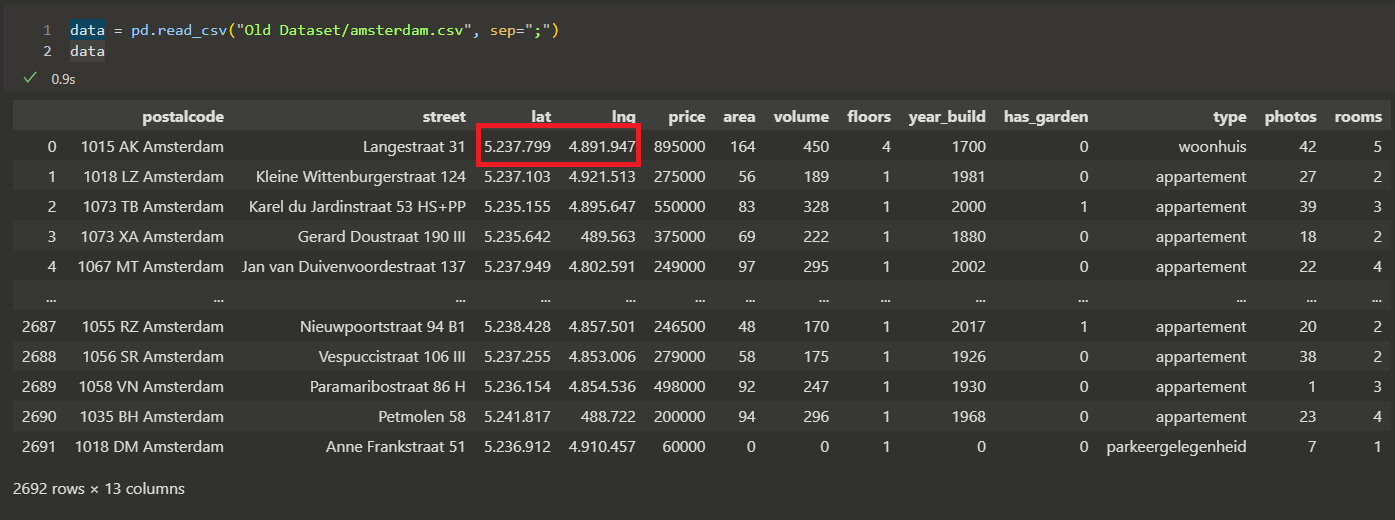
|  |  |
| --- | --- |
| **Data Analysis Final Assignment** | |
|  | |
| Name | Michael Adriel Darmawan |

1. **Business Understanding & Data Understanding**  
     
   Exercise 1.1  
     
   a.   
   For this final assignment, first we need to identify which city district we want to work on. But, from the very beginning, we noticed that there are some noticeable problems with dataset. First of all, there is no single column mentioning city district at all. So, we thought that we can use the latitude and longitude data to find the city district. But it turned out that the latitude and longitude data needed to be fixed as well.  
     
     
   Graphical user interface, text, application

   Description automatically generated  
     
   Therefore, we decided to fix the dataset first by using the following steps:  
   1. Using the postalcode data, find the correct latitude and longitude data (using public API -> https://app.zipcodebase.com/api/v1)  
     
   2. Then, from the correct latitude and longitude data get the address details, including the city district that it belongs to (using https://www.geoapify.com/tools/reverse-geocoding-online).  
     
   In the first step, we created a simple python script that iterates each row of the dataset and creates and API call to get the correct latitude and longitude data.  
   Text

   Description automatically generated   
     
   In the second step, we used the <https://www.geoapify.com/tools/reverse-geocoding-online> website to find the address details, resulting in the dataset that now we can use for our data analysis task.  
   Text

   Description automatically generated  
     
   Upon looking at the data, we can see that the data in West district of Amsterdam has the most data, making it somewhat the most suitable candidate of district to be chosen.  
   Diagram

   Description automatically generated  
   Chart

   Description automatically generated with medium confidence  
   Graphical user interface

   Description automatically generated with medium confidence  
   Graphical user interface, text

   Description automatically generated  
   Table, Excel

   Description automatically generated  
     
   From the pictures above, it can be seen that the West district of Amsterdam has the most data with less outliers, for example when compared to Zuid district which also has a big number of data (334 rows), but it has more outliers data (can be seen in the box plot). We are trying to avoid data that has more outliers because it will make our machine learning model more unreliable. The criteria in choosing the city district here is the one that has the most data with less outliers.  
     
   Moving on, we are trying to determine whether we want to use the apartment type of housing or woonhuis.  
    Graphical user interface

   Description automatically generated with medium confidenceGraphical user interface, application

   Description automatically generated  
     
   By looking at the two pictures, overall, it can be seen that woonhuis has a better data distribution compared to apartment (apartment has more outliers as well). Woonhuis has the median of 877188.17, whereas apartment has the median of 502023.86. So for now the woonhuis type of housing seems like a better option from the two.  
     
     
     
     
     
     
     
     
     
   b.  
   In the city of Amsterdam and Netherlands, WOZ value, which determines a property value, is used to calculate the amount of tax that we owe to the government. The WOZ value for a real estate/property is based on market values, building’s characteristics, official valuations, and nearby real estate selling price.   
     
   c.  
     
     
     
     
     
     
     
     
     
     
   C.   
   First, we choose the city district of West. This is because   
     
     
   Also, we decided early on that we would work only with apartment data. This is because the apartment type of housing has a bigger number of data, meaning that our machine learning algorithm get to work with more data later on, which will resulted in a better training model.
2. **Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla volutpat nec ipsum non tempus.**   
     
   Etiam et quam at metus bibendum convallis vitae quis ligula. Vivamus imperdiet, magna quis pulvinar commodo, nisl dui vulputate ante, sed lacinia libero velit id magna. Quisque eleifend,
3. **Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla volutpat nec ipsum non tempus.**   
     
   Etiam et quam at metus bibendum convallis vitae quis ligula. Vivamus imperdiet, magna quis pulvinar commodo, nisl dui vulputate ante, sed lacinia libero velit id magna. Quisque eleifend, risus ut